CAESAR, LINDSAY K., Ph.D. Bioinformatic Strategies to Understand the Complexities of Medicinal Natural Product Mixtures. (2019)
Directed by Dr. Nadja B. Cech. 299 pp.

Compounds from natural sources, as well as those inspired by them, represent the majority of small molecule drugs on the market today. Plants, owing to their complex biosynthetic pathways, are poised to synthesize diverse secondary metabolites that selectively target biological macromolecules. Despite the vast chemical landscape of botanicals and other natural products, drug discovery programs from these sources have diminished due to the costly and time-consuming nature of standard practices and high rates of compound rediscovery. Additionally, natural product mixtures are incredibly complex, and the standard reductionist approaches often ignore the presence of combination effects such as synergy and antagonism. Bioinformatics tools can be used to integrate biological and chemical datasets, and statistical analyses of these datasets are broadly termed "biochemometrics." Biochemometric approaches enable researchers to predict active constituents early in the fractionation process and to tailor isolation efforts toward the most biologically relevant compounds. Throughout the course of this project, bioinformatics approaches were used to (1) discover biologically active constituents from the botanical medicines, (2) develop and improve data filtering, data transformation, and model simplification parameters to optimize biochemometrics models, and (3) produce a new approach capable of predicting mixture constituents that contribute to synergy, additivity, and antagonism in complex mixtures.

The first goal was achieved by applying bioassay-guided fractionation, biocheomometric selectivity ratio analysis, and molecular networking to

comprehensively evaluate the antimicrobial activity of the botanical *Angelica keiskei* Koidzumi against *Staphylococcus aureus*. This approach enabled the identification of putative active constituents early in the fractionation process, and provided structural information for these compounds. A subset of chalcone analogs were prioritized for isolation, yielding antimicrobial compounds 4-hydroxyderricin, xanthoangelol, and xanthoangelol K. This approach successfully identified a low abundance compound (xanthoangelol K) that has not been previously reported to possess antimicrobial activity.

Two studies were undertaken to achieve the second goal. First ,we demonstrated the effectiveness of hierarchical cluster analysis (HCA) of replicate injections (technical replicates) as a methodology to identify chemical interferents and reduce their contaminating contribution to metabolomics models. Pools of metabolites were prepared from the *A. keiskei* and analyzed in triplicate using ultraperformance liquid chromatography coupled to mass spectrometry (UPLC-MS). Before filtering, HCA failed to cluster replicates in the datasets. To identify contaminant peaks, we developed a filtering process that evaluated the relative peak area variance of each variable within triplicate injections. This filtering process identified 128 ions that did not show consistent peak area from injection to injection that likely originated from the UPLC-MS system. When interferents were removed, replicates clustered in all datasets, highlighting the importance of technical replication in mass spectrometry-based studies and providing tool for evaluating the effectiveness of data filtering prior to statistical analysis.

As a follow up study, the impact of data acquisition and data processing parameters on selectivity ratio models were assessed using an inactive botanical mixture

spiked with known antimicrobial compounds. Selectivity ratio models were used to identify active constituents that were intentionally added to the mixture, as well as an additional antimicrobial compound, randainal, which was masked by the presence of antagonists in the mixture. This study revealed that data processing approaches, particularly data transformation and model simplification tools using a variance cutoff, had significant impacts on the models produced, either masking or enhancing the ability to detect active constituents in samples. This study emphasized the importance of data processing for obtaining reliable information from metabolomics models and demonstrates the strengths and limitations of selectivity ratio analysis to comprehensively assess complex botanical mixtures.

Often, analytical tools aimed to assess biological mixtures ascribe the activity to a few known components. Although researchers recognize this as an oversimplification, research methodologies to address this problem have not been developed. To overcome this and to achieve the third goal of this project, a new approach called Simplify was developed that can both identify mixture components that contribute to biological activity and characterize the nature of their interactions prior to isolation. As a test case, this approach was applied to the botanical *Salvia miltiorrhiza* and successfully utilized to identify both additive and synergistic compounds. These findings illustrate the efficacy of this approach for understanding how natural product mixtures work in concert and are expected to serve as a launching point for the comprehensive evaluation of mixtures in future studies.

BIOINFORMATIC STRATEGIES TO UNDERSTAND THE COMPLEXITIES OF MEDICINAL NATURAL PRODUCT MIXTURES

by

Lindsay K. Caesar

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2019

Approved by

_____
Committee Chair

To Neil B. Caesar, the best dad…
you are always with me. I would not be me without you.

APPROVAL PAGE

This dissertation, written by LINDSAY K. CAESAR, has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

This research was supported by the National Center for Complimentary and Integrative Health under award numbers 1R01 AT006860, 1R15 AT010191, U54 AT008909 (NaPDI, Center of Excellence for Natural Product Drug Interaction Research), and 5T32 AT008938 (Ruth L. Kirschstein National Research Service Award Institutional Research Training Grant).

I would like to thank Drs. Nicholas Oberlies, Qibin Zhang, and Daniel Zurawski for offering their guidance and expertise as committee members throughout this process, Dr. Daniel Todd for training me in mass spectrometry and providing guidance on my projects, Dr. Olav Kvalheim for his vast knowledge of multivariate statistics, Dr. Joshua Kellogg for training in lab, and Sonja Knowles for continuous assistance with NMR interpretation. I especially want to acknowledge the entire Cech lab, including undergraduate students, graduate students, and post-doctoral students, for their continuous support. Thank you Nadja Cech for being such an incredible mentor to me over the last four years. I look forward to many years of fruitful collaboration ahead.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

SYNERGY AND ANTAGONISM IN NATURAL PRODUCT EXTRACTS: WHERE 1+1 DOES NOT EQUAL 2

This chapter has been submitted to the journal Natural Product Reports and is presented in that style. Caesar, L.K., Cech, N.B. Nat. Prod. Rep. Submitted.

Caesar, L.K. wrote all sections of this manuscript except for the section, "Endotoxin from bacterial endophytes in *Echinacea* species" which was written by Cech, N.B. Caesar, L.K. created all of the figures except for Figures 3 and 5-9 which were re-printed (with permission) from existing publications. Ashley Scott made Figure 4. Caesar, L.K. and Cech, N.B. worked together to write the outline for this manuscript and collaborated on the introduction. Cech, N.B. provided suggestions and edits throughout manuscript preparation.

Introduction

Plants have been used as medicine since the beginning of human history (1). Texts from ancient Sumeria, India, Egypt, China, and others contain recipes for medicinal plant preparations for the treatment of disease (1, 2). Today, medicinal plant use remains widespread, and a significant portion of the world's population utilizes herbal natural products and supplements as the primary mode of healthcare (3-5). In the United States, nearly 20% of adults and 5% of children use botanical supplements to treat disease (6, 7).

Despite centuries of use, the activity of botanical medicines is only partially understood, and for most natural products on the market, there is a lack of knowledge as to which constituents are responsible for the purported biological activity. Scientific investigation of botanical natural products is challenging because of their immense

complexity and variability (8-10). Natural products chemistry efforts are typically devoted to reducing complexity and identifying single "active" constituents for drug development. However, given that complex plant extracts, and not single molecules, are often administered for medicinal purposes, interactions between constituents could be of great importance.

Understanding how mixtures work in concert to achieve a given biological effect may address the ever-increasing threat of disease resistance. Indeed, many diseases are not regulated by a single molecular target, but often have a multi-factorial causality (8, 10). It has been shown in numerous studies that disease resistance is less likely to occur against a combination of compounds than to single active constituents (9, 11). Plants have evolved over millennia to address the multifactorial nature of disease pathogenesis by targeting pathogens through the combined action of structurally and functionally diverse constituents (8, 12). As such, complex natural product mixtures offer an important resource for drug development, and to ensure future success in natural products research, understanding interactions within and between the constituents of natural product mixtures is paramount.

Botanical extracts may contain hundreds or even thousands of individual constituents at varying abundance (13) (Figure 1) and identifying the compounds responsible for a given biological effect represents a significant challenge. Too often, it is assumed that the behavior of a mixture can be described by the presence of just a few known constituents. However, a number of studies have shown that the overall activity of botanical extracts can result from mixtures of compounds with synergistic, additive, or

antagonistic activity (10, 14-17), and those who work in the field of botanical natural products research will be quick to admit that it is very often the case that isolation efforts on a botanical extract fail because activity is lost upon fractionation (10, 14, 17). While there are multiple possible explanations for this failure (including irreversible adsorption of compounds to the column packing), it is certainly true that in some cases loss of activity occurs because multiple constituents are required to observe the biological effect. Many investigators recognize the multi-factorial nature of botanical medicines. However, research methodology as applied to botanical mixtures still tends, in most cases, either to take a reductionist approach (focusing on just one or two "marker compounds" or to ignore the question of chemical composition altogether, testing the biological effects of complex mixtures for which active constituents are unknown. The problem in the latter case is that results tend to be difficult both to interpret and to reproduce. Herein, we seek to provide an overview of the methodology that currently exists to understand combination effects within complex mixtures. We will highlight existing technologies for studying combination effects, placing particular emphasis on – Omics technologies and other Big Data approaches that have developed significantly in the last several years. Herein we seek to provide practical advice to investigators seeking to comprehensively evaluate the constituents and mechanisms responsible for the biological activity of botanical mixtures.

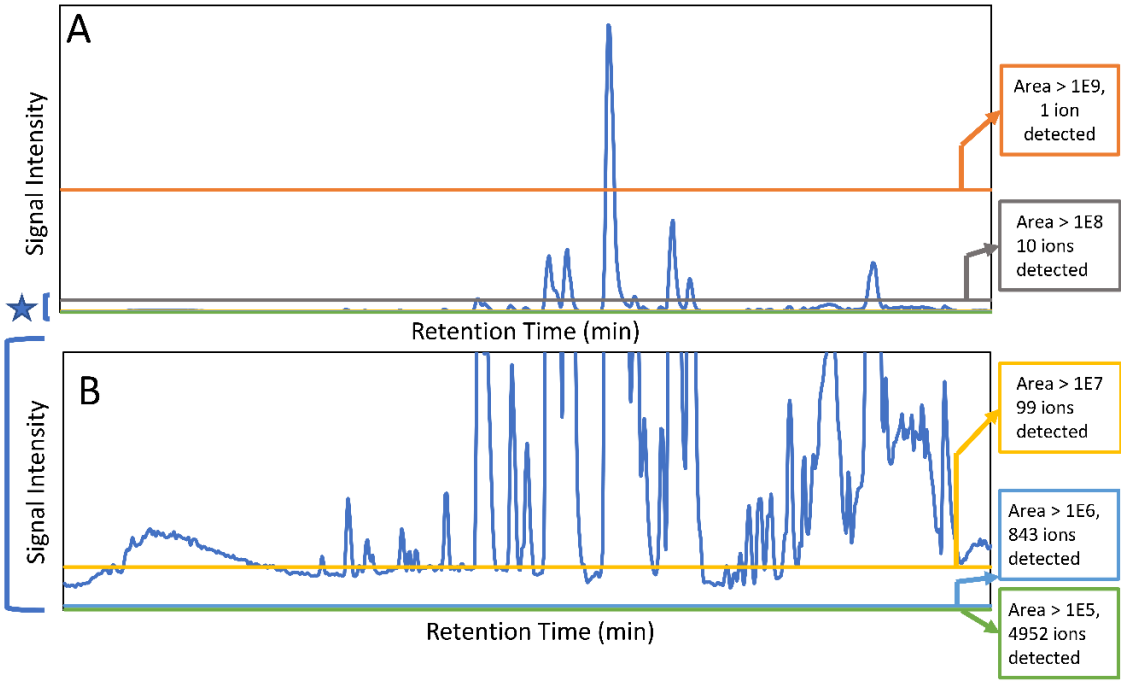


Figure 1. Chromatograms (Obtained with Liquid-Chromatography Coupled to Mass Spectrometry) of a Complex Extract of the Botanical *Salvia miltiorrhiza* (Chinese Red Sage or Danshen). The full chromatogram is shown in (A), while (B) shows a zoomed in version of the baseline that demonstrates the immense complexity of the mixture. Counts for numbers of ions detected are shown at the right, and it is observed that the number of ions detected increases by ~10-fold with each 10-fold decrease in the cutoff for peak area. Notably, each mixture component may be represented by more than one ion, making it difficult to assign specifically the number of mixture components. Nonetheless, the data indicate the immense complexity of the botanical extract.

What are Combination Effects? Definitions of Synergy and Antagonism in the Context of Natural Product Mixtures

Several reviews have been written on the topic of combination effects in recent years that provide valuable commentary on defining combination effects in complex mixtures (8-10). Although the evaluation of interactions between multiple bioactive constituents has gained popularity in many scientific disciplines (14, 18-21), it remains difficult to give a undisputable definition for the term synergy (10, 22). It is generally agreed, however, that interactions between multiple agents can be classified as

antagonistic, additive/non-interactive, or synergistic. Additive and non-interactive combinations indicate that the combined effect of two substances is a pure summation effect, while an antagonistic interaction results in a less than additive effect. Positive interactions, known as potentiation or synergy, occur when the combined effect of constituents is greater than the expected additive effect (8-10, 23-25).

Assays for gathering biological data

To successfully acquire useful data for understanding combination effects in complex mixtures, one must first choose an appropriate biological assay for combination testing. Because combination effects can present themselves through myriad mechanisms (including changes to absorption and metabolism, affecting multiple cell targets, etc.), *in vivo* model systems provide the most comprehensive assessment of the overall effects on a living organism (26). The development of high-throughput *in vivo* testing of mixture-based libraries shows promise for identifying multi-target constituents within mixtures (26). Despite this, it remains challenging to address the complexity of *in vivo* systems, which require the sacrifice of test animals and maintenance of animal facilities. Additionally, results may not successfully translate from one animal model to another. To overcome some of these challenges, many researchers work with *in vitro* systems instead. However, many cell-free, high-throughput assays that search for molecular targets do not accurately model the biology of an intact cell, making the discovery of relevant combination effects unfeasible (27). As such, cell-based assays can be employed that strike a balance between efficiency and preservation of molecular pathway interactions (28). Primary tissue assays comprised of multiple cell types, such as those used to screen

drug combinations for anti-inflammatory activity in mixed cultures of lymphocytes, can also be used to reveal combination effects that work through multi-target mechanisms (27). In addition to carefully choosing the biological system to study for combination effects, data enabling the efficient comparison of a drug combination to agents in isolation must be gathered (9, 27). Combination effects including synergy and antagonism can occur over a broad range of concentrations, so various ratios of the samples under study must be tested (9, 27-32).

Numerous methodologies have been developed to acquire data to discover combination effects *in vitro*, including checkerboard assays and time-kill methods, many of which are quite labor- and material-intensive (9, 27, 33). One of the simplest methods for identifying potential combination effects is through testing samples alone and in combination, and determining if the combined effect of the samples is greater, equal, or less than the expected sum of the two samples in isolation. Although simple, assays employing this approach cannot claim synergy without further study because they lack the range of concentration combinations required to fully assess combination effects, and should be used only to prioritize samples for more in-depth studies (34). These in-depth studies can be achieved using a dose-response matrix design (28), also known as a checkerboard assay, in which a series of dose-response curves using different dose combinations of the agents under study are acquired and compared (9, 27, 28).

In addition to concentration-based approaches to evaluate combination effects, time-based approaches have also been developed and applied to identify antimicrobial synergy and to describe the relationship between bactericidal activity and sample

concentration (35). This method involves exposing a selected pathogen to an inhibitor (or combination of inhibitors), sampling cultures at regular time intervals, serially diluting and incubating aliquots, and comparing the colony forming units produced. The resulting dose response curve can be used to define additive, synergistic, and antagonistic effects (35). Importantly, several of these methods have been compared using the same datasets (29-32), revealing a lack of consistency between conclusions met using these approaches (29-31, 33, 36). Not only do *in vitro* tests often result in conflicting results, but it is very often the case that reproducible hits *in vitro* lack efficacy *in vivo* (26). Because of this inconsistency, preliminary screening efforts should be used to prioritize candidates with *potential* synergy but should not be used to unequivocally define combination effects.

Models for assessing combination effects

To identify if an interaction exists between individual compounds or complex samples, the observed combination response must be compared to the expected effect using a "null reference model"(37, 38). Much of the confusion around categorizing interactions as antagonistic, additive, or synergistic results from the use of different reference models that are used to define the "expected" outcome of a given combination (23, 39-42). As described in a recent paper by Tang et al.(22), the two major reference model classes are the Bliss independence model (43) and the Loewe additivity model (44), each of which relies on a different set of biological assumptions. The Bliss Independence model, for example, assumes that each sample has independent, yet competing effects, while the Loewe Additivity model defines the expected effect as a sample combined with itself (38). Recently, an additional reference model, the Zero

Interaction Potency (ZIP) model, was developed that takes advantage of both Loewe and Bliss models (38). The ZIP model is based on the assumption that two non-interacting samples will cause minimal changes to the dose-response curves, both in terms of the slope of the curve and in the half maximal effect (38). This model shows particular promise for high-throughput drug combination screenings and shows potential for identifying the variety of combination effects that can occur across different concentration ranges (38). These models, and other lesser utilized models, are discussed in depth in several publications (22, 23, 40, 45).

Despite the existence of numerous reference models, the general isobole equation, based on the assumptions of the Loewe Additivity principle, remains the most popular for studying combination effects (9, 10, 23-25, 46). As described elsewhere, an isobole, or an "isobologram," is a graphical representation of the combination effects between two samples (9, 10, 23-25, 46). The axes of the plot represent the doses of individual agents, and the points plotted indicate the combination of concentrations of the two treatments required to reach a particular fixed effect (i.e. 50% inhibition of cell growth) (9). If the two samples have no interaction, the line joining the axes will be a straight line. Synergy will result in a concave curve, and antagonism will result in a convex curve (Figure 2) (9, 10, 23-25). In a recent publication, Lederer et al.(46) scrutinized the implicit assumptions of the Loewe Additivity model (and with it the general isobole equation), and found that the consistency of the model only holds if the two samples under study do not differ in the slopes nor the maximal effects of their dose response curves (46). In cases where one sample reaches an effect that cannot be reached by the other sample, the Loewe

Additivity Consistency Condition is violated (45, 47). To overcome this limitation with the Loewe Additivity Consistency Condition, Lederer et al. (46) developed an adaptation of the general isobole equation termed the Explicit Mean Equation. The Explicit Mean Equation is equivalent to the isobole equation in cases where the two samples meet the Loewe Additivity Consistency Conditions and is capable of identifying combination effects even if this condition is violated. In a follow up study, Lederer et al.(37) compared six models built on either Loewe Additivity or Bliss Independence principles using existing, high-throughput datasets (48, 49) and found that Loewe Additivity models performed better than Bliss Independence at separating synergy relationships from other combination effects, and that the Explicit Mean Equation was the overall best performing model (37).

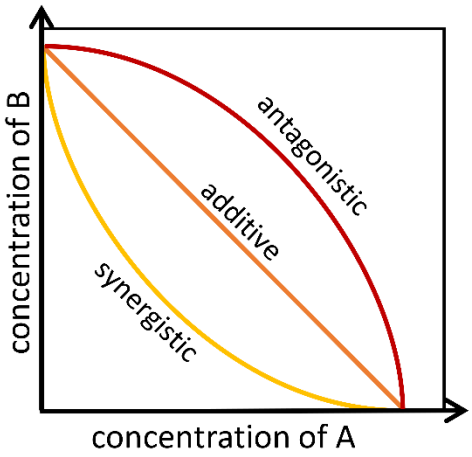


Figure 2. Example of Isobolograms for Antagonistic, Additive, and Synergistic Components. Axes represent the doses of individual agents, and the points represent the combination of concentrations of the two agents required to reach a particular fixed effect.

In recent years, variants of the Loewe additivity model and the Bliss independence model have been developed (47, 50-54). However, because the expected responses from these different models are often disparate (23, 37), it is challenging to draw biological conclusions from the resulting data. In some instances, combination effects have been identified as synergistic by one model but antagonistic by another (48). As such, researchers should be clear about which model they have chosen to adopt, as stated in the Saariselkä agreement (55). Tang et al.(22) have expanded upon this suggestion and have proposed the use of terminology that incorporates results from both Bliss Independence and Loewe Additivity models. Tang et al.(22) argue that the level of consistency between models should be used to designate the degree of synergy or antagonism. For example, if both models identify a given interaction as synergistic, that interaction should be considered "strong synergy," and if the combination is identified as synergistic by one model only, it should be considered "weak synergy"(22). By utilizing both models, this proposal minimizes the incorporation of bias into predictions and provides more informative definitions for the combination effects described. While in principle this proposal makes sense, it also relies on the assumption that the models are equally valid. While Loewe Additivity models have been shown to perform better than Bliss Independence models on numerous occasions (37), Russ and Kishony (56) found that the Bliss Independence models are more consistent when interactions between more than two samples are evaluated. As such, the use of any synergy model should be seen only as a hypothesis-generating tool to prioritize potential interaction effects for further study. Indisputable definitions of synergy and antagonism remain elusive, and a wider

agreement on the terminology used for interaction assessment is still required to standardize future research initiatives.

Scoring and interpreting biological data

In addition to a lack of consensus among the theoretical models to utilize for defining combination effects, there are challenges on how to apply and interpret existing models to analyze drug combinations (38). As stated earlier, most synergy analyses focus on the differences in isobologram shapes at fixed effects, and summary interaction scores such as the fractional inhibitory concentration (ΣFIC) index have found wide application (9, 38, 39, 57, 58). The ΣFIC index is calculated using equation 1(9):

$$\Sigma FIC = FIC_A + FIC_B,$$

$$\text{Where } FIC_A = [A]/IC50_A, \text{ and } FIC_B = [B]/IC50_B \quad \text{(equation 1)}$$

In this equation, A and B represent the samples under study, $IC50_A$ and $IC50_B$ represent the concentrations of A or B in isolation to reach 50% inhibition, [A] is the IC_{50} of A in the presence of B, and [B] is the IC_{50} of B in the presence of A. Notably, any fixed effect can be used to calculate ΣFIC indices, but IC_{50} values are perhaps the most common metric.

Despite the popularity of this method, the interpretation of ΣFIC scores for defining combination effects varies considerably from author to author. In their recent publication, van Vuuren and Viljoen provide an excellent commentary on ΣFIC score interpretation (9). The earliest interpretations by Berenbaum considered synergistic interactions to be any value below one, additive/indifferent interactions focused on one,

and antagonistic interactions above one (23). However, because of the inconsistency across null reference models, and because fixed effects can often be placed within a three-dilution range using *in vitro* assays (59), a more conservative approach is warranted. Taking this into consideration, van Vuuren and Viljoen (9) and the authors of this review suggest that synergistic interactions be defined as interactions having $\Sigma FIC \leq 0.5$, additive interactions range from 0.5 to 1.0, non-interactive effects range from 1.0 to 4.0, and antagonistic effects fall at or above 4.0 (Table 1).

Table 1. Recommended Fractional Inhibitory Concentration (ΣFIC) Indices for Assigning Combination Effects.

Combination Effect	ΣFIC range
Synergy	$\Sigma FIC \leq 0.5$
Additivity	$0.5 < \Sigma FIC \leq 1.0$
Indifference	$1.0 < \Sigma FIC \leq 4.0$
Antagonism	$4.0 < \Sigma FIC$

Despite its popularity, the ΣFIC index, like the isobologram upon which it is based, is insufficient to effectively capture the combination effects that may occur across multiple dose regions (37, 38). An inherent limitation of the ΣFIC index is the focus on a single interaction parameter. In a recent publication, Lederer et al. (37) compared multiple synergy measurements and found that the "lack of fit" model (60), where synergy scores are defined by the volume spanned between the null reference model and the measured response, performed better than parametric models in its ability to identify synergistic effects (37).

Similarly, Yadav et al. developed a score that enables the use of an interaction landscape over the full dose-response matrix to identify combination effects across

multiple dosages and response levels (38). Rather than relying on a single parameter such as the IC_{50} measurement, the delta-score utilized in this study was calculated by assessing changes in both the shape parameter and the midpoint of each dose response curve for individual samples and combinations thereof. The delta scores were visualized using a response surface plot to visualize the combination effect landscape over all tested dosage combinations, enabling identification of potency changes and differences in combination effects even within the same sample pair (Figure 3).

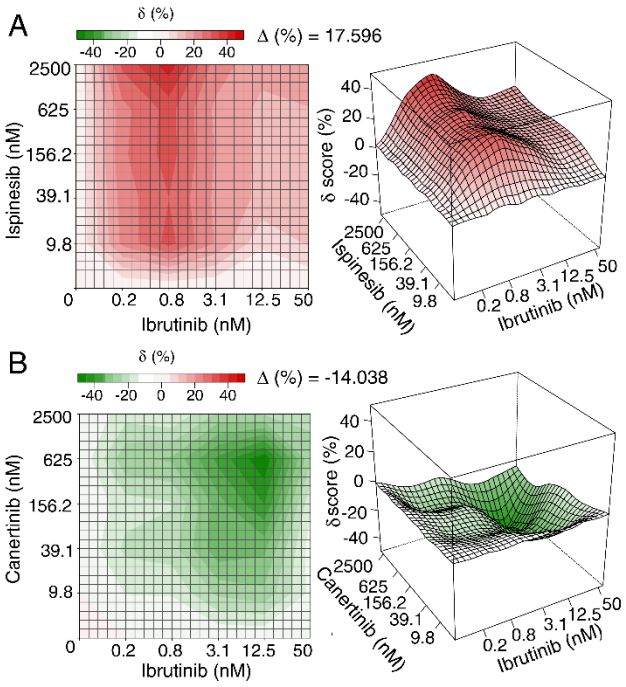


Figure 3. Example of Synergistic (Top) and Antagonistic (Bottom) Interaction Landscapes using Delta Scores (δ) Calculated with the Zero Interaction Potency Model of Compounds in Combination with Ibrutinib, an Approved Anti-Cancer Drug Targeting Bruton's Tyrosine Kinase. (A) Interaction map between anti-cancer activity of ispinesib (a selective kinesin spindle protein inhibitor) and ibrutinib. Average delta across the dose response matrix (Δ) is 17.596, indicative of overall synergy. (B) Interaction map between canertinib (an epidermal growth factor receptor family inhibitor) and ibrutinib. The Δ value is -14.038, indicative of overall antagonism. Figure is reprinted with permission from Yadav et al. 2015 (38).

There appears to be value in using these different methods to explore synergy; however, these approaches have not yet been applied to understand synergy in complex natural products and discussion of their merit for this purpose remains hypothetical. Despite the aforementioned limitations, isobole analysis and the ΣFIC index have found the widest utility in natural products research (9).

Documented Examples of Natural Products that Contain Synergists or Antagonists

Proponents of the health benefits of plant-based medicines often proclaim that whole plant preparations are more effective than isolated compounds due to the beneficial interactions between constituents within them (10, 15, 61, 62). While this claim is sometimes disputed (63-66), considerable evidence exists that combination effects within complex extracts can alter the biological activity of a mixture (8-10, 67). Here, we provide a few case studies in which synergy and/or antagonism within botanical preparations have been discussed. Additional examples of synergy within and between botanical extracts have been extensively reviewed in several publications (9, 10, 17, 24, 67), providing compelling evidence that at least in some cases, the combined effect of botanical mixtures is not simply a summation of their individual constituents. However, explorations into phytosynergy are only in their infancy. The vast majority of complex natural product mixtures still await chemical investigation, representing an untapped resource with considerable potential for future scientific exploration.

Anti-plasmodium activity of *Artemisia annua*

Artemisia annua L. (Asteraceae) has gained considerable popularity over the last few years since the award of the 2015 Nobel Prize in Physiology or Medicine to Youyou

Tu for her discovery of artemisinin, an antimalarial sesquiterpene lactone produced by this plant (68, 69). Artemisinins have been established as potent and safe antimalarial agents (70), and artemisinin-based combination therapies are now the front-line treatment recommendation by the World Health Organization (71). The replacement of ineffective malaria treatments such as chloroquine with artemisinin-based combination therapies has decreased malaria-associated morbidity and mortality worldwide (72-74). Several researchers have suggested that artemisinin acts to destroy *Plasmodium falciparum* parasites through the activation of a trioxane bridge in the *P. falciparum* food vacuole in a heme-dependent manner (75, 76). This disruption causes the production of free radicals that interrupt heme detoxification, ultimately generating more reactive oxygen species and killing the parasite.

In addition to artemisinin, there are approximately 30 additional flavonoids and sesquiterpenes within *A. annua*, some of which have minor anti-plasmodial activities (77). As might be expected, since botanical preparations are multi-factorial rather than monospecific in nature, both *in vitro* and *in vivo* studies evaluating the activity of *A. annua* extracts have found that the amount of artemisinin in the extracts does not fully explain the extract's efficacy against *P. falciparum* parasites (78, 79). Indeed, various combination therapies including artemisinin and its derivatives are utilized as antimalarial treatments (80, 81). In a recent study, Suberu et al. (82) aimed to identify the compounds within *A. annua* tea extract contributing to its anti-plasmodial efficacy. Building upon the work of previous studies which found that several flavonoids potentiated the activity of artemisinin against *P. falciparum* (83, 84), Suberu et al. (82) tested the tea extract,

purified compounds from the extract, and various combinations of artemisinin with purified compounds against both chloroquine-sensitive and chloroquine-resistant strains of *P. falciparum.* Interestingly, the type of combination effect observed, whether it be synergistic, additive, or antagonistic, often differed depending on the dosage of the combined constituents and/or the resistance profile of the parasite under analysis (82).

Using isobologram analysis and calculating ΣFIC indices, Suberu et al. (82) found several compounds that enhanced or antagonized the activity of artemisinin against *P. falciparum.* Two compounds that contained anti-plasmodial activity, 9-epi-artemisinin and artemisitene, were found to antagonize the efficacy of artemisinin against both chloroquine-sensitive and chloroquine-resistant strains. Although the mechanism by which these compounds antagonize artemisinin's activity is unknown, it is reasonable to assume these compounds, which have only minor structural differences, compete for the same molecular target, reducing the overall efficacy of the compounds in combination (82). Several additional compounds contained within the extract, however, did not demonstrate the same combination effect at all concentrations tested. For example, 3-caffeoylquinic acid showed a summation effect in combination with artemisinin at a ratio of 1:3 (artemisinin to 3-caffeoylquinic acid) when tested against the chloroquine-sensitive strain, but at higher combination ratios (1:10-100), synergistic interactions were observed. Similarly, casticin, which possessed antagonistic activity at the 1:3 ratio, has been reported to be synergistic in other studies using higher combination ratios (1:10-1000) (83, 84). The reason for this discrepancy is unknown, but it is possible that these compounds act as either anti-oxidant or pro-oxidant species depending on the dosage

level (85, 86). When combined at a low concentration with artemisinin, they may have counteracted artemisinin activity through anti-oxidative interaction, minimizing the oxidative stress resulting from the reactive oxygen species formed through artemisinin's activity, while at higher concentrations they were pro-oxidative, increasing the oxidative stress and leading to increased efficacy of artemisinin (82).

Other compounds, including rosmarinic acid and arteannuin B, showed differential combination effects when tested against sensitive and resistant strains of *P. falciparum.* Rosmarinic acid was synergistic against the sensitive strain, but showed antagonistic activity in the resistant strain (82). Similarly, arteannuin B had an additive/indifferent interaction in the chloroquine sensitive strain, but a synergistic interaction with the resistant strain, leading to a three-fold improvement in artemisinin's activity (82). Because arteannuin B selectively potentiates the activity of artemisinin in the chloroquine-resistant strain, it likely targets the parasite's chloroquine resistance mechanism, illustrating the promise of combination treatments not only for developing therapeutics against drug-resistant pathogens, but also for providing insight into the mechanisms by which parasites gain resistance as a whole.

It is important to note that Suberu et al. chose somewhat liberal ranges for the ΣFIC indices used to define their combination effects (82), and other researchers, depending on the models chosen, may have categorized some of the synergistic and antagonistic interactions as "additive" or "indifferent"(9). Even if one were to re-categorize interactions based on conservative estimates, however, all three types of combination effects (synergy, additivity, and antagonism) were witnessed during the

course of this study. While the specific categorizations of synergy, additivity, and antagonism chosen by Suberu et al. may be disputed, it is clear that the nature of combination effects did often change depending on both the dosage and the parasite strain under study (82).

Endotoxin from bacterial endophytes in *Echinacea* species

Few botanicals have been the subject of as much research or as much controversy as plants from the genus *Echinacea*. This botanical, which is widely used for the treatment of upper respiratory infections, has been the subject of several clinical trials. Although these trials had conflicting results (87, 88), *Echinacea* species remain one of the most popular and best-selling botanical medicines in the United States (89), and preparations from this plant are popular in Europe as well (90).

The constituents responsible for the activity of *Echinacea purpurea* extracts and the mechanisms by which these constituents exert their purported beneficial effects have been studied extensively. Early research on *Echinacea* attributed its purported health benefits to its ability to "activate" or "stimulate" immune cells. These findings were based upon early work by Wagner and co-workers, in which isolated *Echinacea* polysaccharides were observed to stimulate phagocytosis and induce TNF-alpha secretion by macrophages (91). Later research demonstrated that much of the immunostimulatory activity originally attributed to *Echinacea* polysaccharides could instead be linked to the lipopolysaccharides and lipoproteins. These lipoproteins and lipopolysaccharides are components of bacterial cell walls, and can be attributed to the presence of bacterial endophytes, bacteria living asymptomatically within the *Echinacea* plant tissues (92-95).

Even very minute quantities of certain lipoproteins and lipopolysaccharides induce pronounced immunostimulatory effects in macrophages, so the presence of these compounds as contaminants can confound *in vitro* assay data.

An alternative narrative about the immunomodulatory activity of *Echinacea* preparations focused on alkylamide constituents. Contrary to the research on polysaccharides, lipoproteins, and lipopolysaccharides, these alkylamides were observed to suppress the production of pro-inflammatory cytokines by macrophages (96-99). Such activity could translate to a beneficial anti-inflammatory effect *in vivo*. The apparently contradictory activity of various classes of compounds, both isolated from *Echinacea*, suggested the possibility that the activity of some constituents might be masked by others in the context of complex *Echinacea* extracts. This was shown in a study by Todd et al. (95), in which complex *E. purpurea* extracts possessed little to no cytokine-suppressive activity, but could be separated to produce sub-fractions with opposing effects. Some fractions, those containing alkylamides, suppressed cytokine and chemokine production by macrophages, while others, those containing lipopolysaccharides, induced cytokine production. Thus, it was demonstrated that lipopolysaccharides (and likely other compounds of bacterial origin) masked the anti-inflammatory effect of complex *Echinacea* preparations, effectively acting as antagonists. It was not until these fractions were separated that the individual activities of the various constituents could be observed.

Underlying Mechanisms of Synergy

Synergy can occur through a variety of mechanisms, including (i) pharmacodynamic synergism through multi-target effects, (ii) pharmacokinetic synergism

through modulation of drug transport, permeation, and bioavailability, (iii) elimination of adverse effects, and (iv) targeting disease resistance mechanisms (9, 10, 67, 68, 100). While the general mechanisms by which synergy can occur are relatively well studied, the mechanisms by which specific botanical preparations exert synergistic effects remain largely unknown (67, 101), stymying efforts to standardize and optimize them for therapeutic purposes. Only through understanding the nature of synergistic activity within botanical extracts will we be able to optimize safe and efficacious preparations for the treatment of disease.

Pharmacodynamic synergism

Cancerous cells and pathogenic organisms can quickly gain resistance to drugs containing a single compound, and many cancers and resistant bacterial infections are now treated with complex drug combinations affecting multiple targets to overcome the development of resistance (102, 103). Plants have long had to defend themselves against multi-factorial diseases, and have evolved to produce multiple active constituents that can adhere to cell membranes, intercalate into RNA or DNA, and bind to numerous proteins (8, 104-106). Pharmacodynamic synergism results from the targeting of multiple pathways, which can include enzymes, substrates, metabolites, ion channels, ribosomes, and signal cascades (10, 107).

Oftentimes, disease targets are able to counteract the therapeutic effect of an active metabolite, resulting in its reduced efficacy (67). One type of pharmacodynamic synergism involves "anti-counteractive action" in which a synergistic compound binds to an anti-target, effectively inhibiting the disease target from counteracting the therapeutic

effect of the active constituent (67). Pharmacodynamic synergy may also occur through complementary actions, in which synergists in a mixture interact with multiple points of a given pathway, resulting in positive regulation of a process affecting the drug target or in the negative regulation of competing mechanisms. Through the selective variation of target activity and expression through complementary actions, pharmacodynamic synergists can both augment beneficial effects of treatments and reduce adverse effects of the disease (67). For example, *Ginkgo biloba* has been shown in numerous studies to have synergistic neuroprotective effects both *in vivo* and *in vitro* by inhibiting the formation of free radicals, scavenging reactive oxygen species, regulating gene expression of mitochondrial targets, and reducing excessive stimulation of nerve cells by neurotransmitters (57, 108).

Pharmacokinetic synergism

In addition to pharmacodynamic synergy, plants often contain compounds that do not possess specific pharmacological effects themselves, but increase the solubility, absorption, distribution, or metabolism of active constituents (8, 10, 67, 109). These pharmacokinetic effects result in enhanced bioavailability of active constituents, enabling increased efficacy of the extract as compared to individual constituents in isolation (10). Several examples exist in which mixture constituents improve the solubility of active constituents. For example, hypericin from Saint John's Wort (*Hypericum perforatum),* is poorly soluble in water. However, when hypericin is combined with *H. perforatum* mixture constituents including procyanidin B2 and hyperocide, solubility and oral bioavailability of hypericin are significantly improved (110). Absorption of active

constituents can be improved through a variety of mechanisms, including the inhibition of drug exporters such as P-glycoproteins (100, 111, 112). Additionally, transport barriers may be disrupted or their recovery delayed, improving permeability of active constituents into target cells (67). For example, the absorption of baicalin, a constituent of the plant *Scutellaria baicalensis*, is synergistically enhanced by the addition of both coumarins and volatile oils from the botanical *Angelicae dahurica*, likely by affecting transport systems independent of P-glycoproteins (113). Pharmacokinetic synergy also results from constituents that inhibit enzymes that convert drugs into excretable or inactive forms, or that activate enzymes that convert pro-drugs into active forms (8, 67).

Elimination of adverse effects

An additional type of synergy occurs when inactive mixture constituents serve to neutralize the unwanted side effects of a toxic, yet bioactive constituent. This type of synergy, if it can truly be called that, does not function to improve the efficacy of the active compound(s) *per se,* but rather functions to minimize the negative effects that the active agent may cause (10). Many potent chemotherapeutic agents, for example, while successful in targeting tumor cells, are often limited by severe side effects caused by action of active agents against healthy cells. In a recent study, an extract of staghorn sumac (*Rhus hirta*) was combined with the chemotherapeutic drug 5-fluorouracil (5-FU) commonly used to treat breast and colon cancer (114). In combination with 5-FU, the *R. hirta* extract was found to protect normal cells from 5-FU toxicity *in vitro.* This chemoprotective effect may have be attributed in part to the presence of antioxidants in

the *R. hirta* extract (115), which minimized oxidative stress and cell damage initiated by 5-FU treatment (114).

Targeting disease resistance mechanisms

Many diseases, such as cancers and infectious diseases, have evolved resistance to single-target drugs. In cancer, drug resistance to single chemotherapeutic agents has increased largely due enzymatic cross-talk (116) and counteractive pathways (117, 118). Combination chemotherapy is growing in popularity, in part due to the ability for multi-constituent mixtures to modulate different pathways and overcome drug resistance (119). Infectious diseases, including those caused by fungi (120), viruses (121), and bacteria (122), are also becoming more challenging to treat due to the development of drug resistance. For example, bacterial pathogens gain resistance to antibiotics due to three major reasons: (i) active site modification resulting in inefficient drug binding, (ii) metabolism of antibiotics into inactive forms, or (iii) efflux of antibiotics out of bacterial cells (Figure 4) (17, 100). By targeting these resistance mechanisms, it may be possible to re-sensitize resistant organisms to existing treatments and to slow the development of resistance.

Many bacteria have gained resistance to beta-lactam antibiotics such as penicillin and ampicillin due to the development of beta-lactamase enzymes that cleave the antibiotics into inactive forms (123). One strategy for synergistically overcoming this resistance mechanism is to combine beta-lactam antibiotics with beta-lactamase inhibitors. In a recent study, Catteau et al. (124) found that a dichloromethane extract of shea butter tree leaves (*Vitellaria paradoxa*) synergized the activity of ampicillin,

oxacillin, and nafcillin against methicillin-resistant *Staphylococcus aureus* by targeting PBP2a +/- beta-lactamase enzymes. Ursolic acid and oleanolic acid, major constituents of *V. paradoxa*, were found to be responsible both for synergistic enhancement of beta-lactam activity and also possessed antimicrobial activity of their own (124).

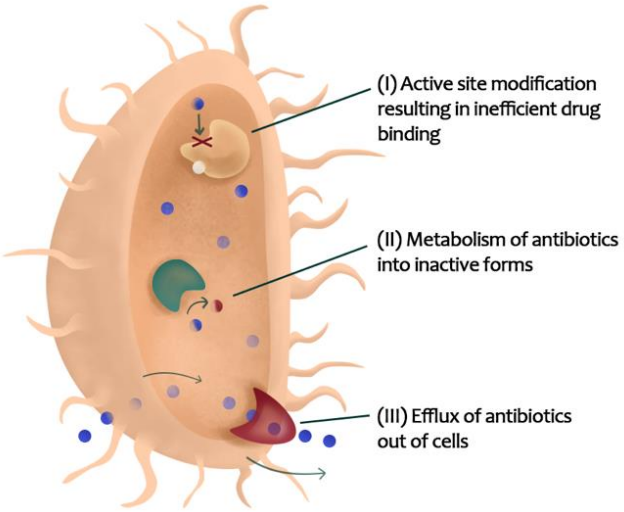


Figure 4. Bacterial Resistance Mechanisms that Could be Targeted with Combination Therapy Enabling Re-Sensitization of Resistant Organisms to Existing Antibiotics.

Another resistance mechanism that protects microbes from antimicrobial agents is the presence of promiscuous efflux pumps that extrude a wide array of compounds from bacterial cells (125, 126). While many naturally-occurring compounds, particularly positively charged alkaloids, are efflux pump substrates, many medicinal plants also contain inhibitors that target efflux pumps, potentiating the effects of antimicrobial agents. In their hallmark paper, Stermitz et al. (15) described the presence of an inhibitor of the norA efflux pump, 5'-methoxyhydnocarpin, in *Berberis* species that potentiated the activity of the antimicrobial compound berberine. More recently, the berberine-

containing plant *H. canadensis* was found to contain norA efflux pump inhibitory activity (127). Many of these synergistic efflux pump inhibitors have been characterized in subsequent publications (14, 18, 128).

Identifying Constituents Responsible for Combination Effects

When working with complex natural product mixtures, constituents responsible for activity are often not known. Additionally, the composition of natural product extracts varies depending on how and where the source material is grown, prepared, processed, and stored (129), and as such, there is a lack of knowledge for many natural products about the dosage and identity of what is being consumed. To address this safety risk, and to improve efficacy of natural product mixtures, bioactive mixtures should be comprehensively characterized and the concentrations and identities of constituents contributing to the biological activity (whether it be through additive, synergistic, or antagonistic means) should be determined. This task, while straightforward in theory, is quite challenging in practice since the biologically important constituents are often not known and are part of a complex matrix containing hundreds or thousands of unique constituents (13).

Isolation and structure elucidation

Bioactivity-guided approaches to identify active molecules

One of the most common approaches for identifying bioactive mixture components is bioassay-guided fractionation. With this approach, active extracts are separated using a variety of chromatographic techniques, the simplified fractions screened for biological activity, and the process iteratively repeated until active

compounds have been isolated and characterized (14, 130-134). In the last decade, substantial developments have been made in improving extraction and separation efficiency, and facilitating the isolation of minor constituents that may contribute to activity (130). Despite the historical effectiveness of bioassay-guided fractionation (135), loss of activity during fractionation is very common (131, 134). Additionally, because structural information is not used to guide separations, this approach may result in the repeated isolation of previously described molecules (134).

To avoid re-isolation of known molecules, preliminary structural assessment steps to identify and discard samples containing known active constituents can be taken (134, 136-138). This process, termed "dereplication," enables prioritization of samples likely to contain new biologically important entities, facilitating efficient use of resources for compound discovery (134, 136-138). Dereplication is often achieved by comparing the spectral patterns of mixture constituents through mass spectrometry (136, 139-141), NMR (142), or UV spectroscopy (136), and searching for known compounds with matching spectral fingerprints in a dereplication database. Recently, the Global Natural Product Social molecular networking (GNPS) platform has been developed that enables spectral annotation and identification of related compounds using MS/MS molecular networking (134, 138, 143). In addition, GNPS provides researchers the ability to share raw MS/MS spectra online, enabling crowdsourced spectra annotation and knowledge sharing between laboratories around the world (138).

Bioactivity-guided approaches to identify synergists

While dereplication protocols have advanced significantly, reducing the likelihood of compound rediscovery, bioassay-guided fractionation may be unsuitable for identifying synergistic compounds from complex mixtures (14, 131). Often, synergistic compounds possess no biological activity on their own, but enhance the activity of active compounds in combination (23). If these compounds are separated from active compounds during the fractionation process, they may be overlooked. Recently, a modification of bioassay-guided fraction was developed, termed "synergy-directed fractionation," which combines chromatographic separation and synergy testing in combination with a known active constituent in the original extract (14). With this process, extracts are subjected to synergy testing, active extracts are fractionated, and resulting fractions again tested for synergy. This process is repeated iteratively until pure compounds have been obtained (Figure 5). By combining fractions with a known active constituent and testing for combination effects, synergists that did not possess activity on their own could still be identified. This approach enabled the identification of three synergists in the botanical medicine *Hydrastis canadensis* that potentiated the activity of berberine through NorA multidrug resistance pump inhibition that would have been overlooked using conventional techniques (14).

Metabolomics and biochemometrics

Metabolomics approaches to identify active constituents

While bioassay-guided fractionation (and modifications of it such as synergy-directed fractionation) have improved significantly with advancements in separation

techniques and dereplication protocols, these methods have a tendency to focus toward the compounds that are most easily isolated in the mixture rather than those that are most likely to be active (18, 132). Thus, it would be desirable instead to identify bioactive compounds in complex mixtures before several rounds of bioactivity-guided fractionation steps have been completed. Towards this goal, many researchers have sought to guide isolation efforts by combining chemical and biological profiles of samples under analysis to identify markers of activity (131-134, 144-147). Using approaches broadly termed as "biochemometrics," chemical and biological datasets can be interpreted using multivariate statistics and putative bioactive constituents identified early in the fractionation process (131-134, 144-147). Biochemometrics has been successfully employed by several research groups to identify minor active constituents from complex natural product mixtures. For example, in a recent study assessing the anti-tuberculosis activity of the Alaskan botanical *Oplopanax horridus*, a total of 29 bioactive constituents were identified based on biological and gas chromatography-mass spectrometry data. Importantly, nearly half of the bioactive constituents identified (14 out of 29) had individual peak areas accounting for less than 1% of the active fraction chromatograms (131).

In mass spectrometry-based biochemometrics studies, the number of variables (ions detected) tends to greatly outnumber the number of samples analyzed (i.e. extracts or simplified fractions), posing a problem for many multiple regression models (148). Partial least-squares (PLS) analysis, however, due to its combination of principal component analysis (PCA) and multiple regression analysis, is less affected by this

mismatch between sample and variable number and is the most popular tool for modelling biochemometric data (148). The resulting PLS models, however, are often incredibly complex and difficult to decipher. Numerous data visualization tools have been developed to extract meaningful information from PLS datasets (132, 148-150).

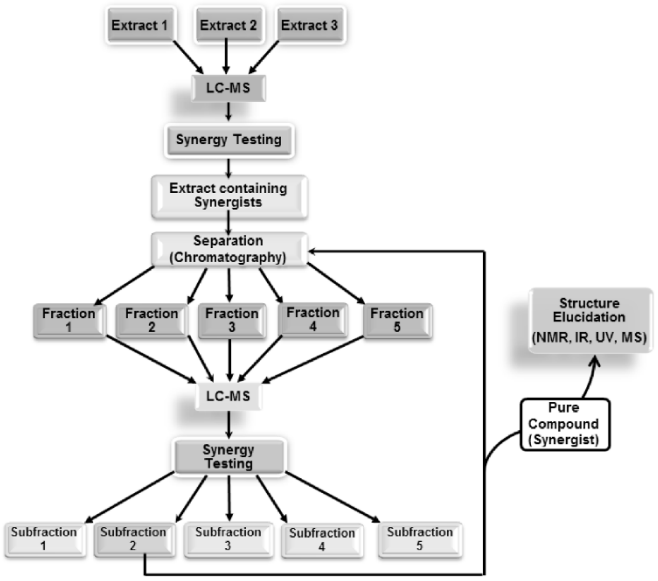


Figure 5. Synergy-Directed Fractionation Workflow. Reproduced with permission from Junio et al. 2011 (14).

One commonly used tool is the S-plot, in which correlation and covariance of variables with a given biological activity are plotted. In a recent study, S-plots were utilized to identify differences in metabolite profiles (detected using UPLC-QTOF-MS) of *Garcinia oblongifolia* leaves, branches, and fruits and to correlate those differences to differences in biological activity (147). Using this approach, 12 marker compounds were identified, primarily xanthones, that were likely responsible for the enhanced antioxidant and cytotoxic properties of the branch extract over other plant parts (147). In another

study, S-plots were generated from bioactivity and chemical profiles of *Ganoderma sinense* to identify potential anti-tumor agents. This approach successfully identified five known cytotoxic compounds with significant antitumor potential (146). A recent study compared the use of S-plot analysis with an additional data visualization tool, the selectivity ratio, to identify antimicrobial constituents from the fungal organisms *Alternaria* and *Pyrenochaeta* sp.(132). In this study, both S-plot and selectivity ratio analyses identified macrosphelide A as the dominant bioactive constituent from *Pyrenochaeta* sp. However, when attempting to identify bioactive compounds from *Alternaria* sp., the selectivity ratio outperformed the S-plot in its ability to identify altersetin, a low abundance antimicrobial constituent, without being confounded by highly abundant (and only weakly active) constituents in the mixture (132).

In a follow up study, an inactive mixture was spiked with known antimicrobial compounds to identify the impact of data acquisition and data processing parameters on biochemometric analysis using the selectivity ratio plot (144). This study found that data transformation, contaminant filtering, and model simplification tools had major impacts on the selectivity ratio models, emphasizing the importance of proper data processing approaches for extracting reliable information from biochemometric datasets (144). In all selectivity ratio studies applied to identify bioactive natural products (132, 144, 145), bioactive mixture constituents were identified early in the fractionation process, enabling chromatographic isolation efforts to be tailored towards mixture constituents that were most likely to possess biological activity.

These numerous examples illustrate the efficacy of biochemometrics for distinguishing between active and inactive chemical entities in complex mixtures. However, these approaches do not provide structural information about putative unknown active constituents, hindering the ability to truly optimize isolation efforts. To address this gap, a recent study utilized a combination of selectivity ratio analysis and GNPS molecular networking to identify putative active constituents from the botanical medicine *Angelica keiskei* and the molecular families to which they belonged (145). Using this approach, a subset of chalcone analogs were targeted for isolation, yielding two known antimicrobial constituents and an additional, low-abundance compound not previously known to possess antimicrobial activity (145). This concept was streamlined into a process called "bioactive molecular networking," in which bioactivity predictions are directly visualized in molecular networks themselves, where the size of individual nodes correspond to the predicted bioactivity score for each ion (Figure 6) (134).

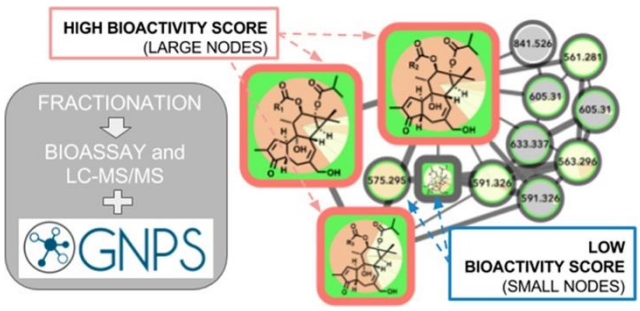


Figure 6. Bioactive Molecular Networking in which Nodes Connected in a Network Represent Structurally Related Compounds Based on MS/MS Fragmentation Patterns, and the Size of Nodes Represents the Correlation of Compound Peak Areas with Biological Activity of Interest. Figure is reprinted with permission from Nothias et al. 2018 (134).

By including both MS/MS fragmentation data and peak area data in the production of molecular networks, bioactive molecular networking enables dereplication, compound annotation, and identification of putative active compounds in one step (134).

An additional approach, Compound Activity Mapping was developed by the Linington laboratory that utilizes image-based cytological screening data and high-resolution mass spectrometry-based metabolomics data to predict both the identities and biological functions of putative bioactive constituents early in the fractionation workflow (133). Using Compound Activity Mapping, biological and chemical datasets are integrated to identify putative bioactive constituents that show consistent positive correlation with phenotypes of interest (Figure 7) (133). The data are presented as a network display, enabling identification and prioritization of lead compounds, even those of low abundance, that likely contribute to a specific biological activity.

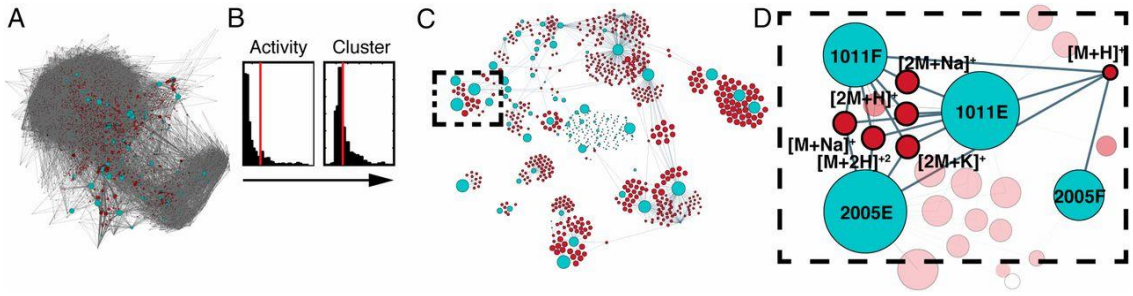


Figure 7. Compound Activity Mapping Workflow. (A) Network analysis of the full chemical space of the tested actinobacterial extracts. Light blue nodes represent extracts connected to all *m/z* features (red), illustrating the immense chemical complexity of the extract library. (B). Activity histograms and cluster scores for all *m/z* features. (C) Compound Activity Map, displaying only extracts and *m/z* features predicted to be responsible for consistent phenotypes of interest. (D) Close up of a specific bioactive cluster, belonging to the staurosporine natural product family. This figure is reprinted with permission from Kurita et al. 2015 (133).

The utility of this approach was demonstrated through the investigation of 234 extracts of actinobacterial origin (133). Using Compound Activity Mapping, biological and chemical datasets from these samples were combined to identify 13 clusters of bioactive fractions containing 11 known molecular families and four new compounds. Subsequent isolation efforts targeted towards these new compounds revealed the presence of a new natural product family, the quinocinnolinomycins, which were predicted to elicit a cytotoxic response through the induction of endoplasmic reticulum stress (133).

Metabolomics approaches to identify synergists

In a recent study, an inactive botanical extract was spiked with four known antimicrobial compounds to assess the ability of selectivity ratio analysis to identify known constituents. Despite the fact that the spiked extract contained concentrations of active constituents that should have completely inhibited the growth of *Staphylococcus aureus*, the extract only caused about a 30% reduction in growth even at the highest concentration tested. To assess the large discrepancy between the predicted and observed activities of the spiked extract, checkerboard assays were conducted in combination with the active constituents berberine and magnolol, yielding ΣFIC indices of 3 and 5, respectively, and strongly indicating the presence of antagonists in the mixture. After chromatographic separation had been conducted, however, antagonists were separated from active constituents and activity of the mixture was restored (144). In a traditional natural products discovery setting, this extract may not have been targeted for isolation efforts despite the fact that it contained active compounds. This example illustrates that predictive tools capable of identifying active compounds alone may not be sufficient to

comprehensively model the complexity of natural product mixtures, and approaches capable of identifying the presence of synergists and antagonists that may not possess any biological activity on their own are needed.

To identify synergists and additives in complex botanical mixtures, Britton et al. recently combined biochemometric analysis with synergy-directed fractionation to identify mixture components from *Hydrastis canadensis* that enhanced the antimicrobial efficacy of berberine through additive or synergistic mechanisms (18). In this study, mass-spectrometry datasets were combined with biological assay data to produce selectivity ratio plots predicting putative additives and synergists (Figure 8). In these plots, negative selectivity ratios are indicative of biological activity, because growth inhibition data (where smaller values indicate biological activity) were used to guide the models. Unlike other biochemometric studies of its kind (132, 145), the biological activity data used in this study did not measure of antimicrobial activity, *per se*, but was rather measured each sample's ability to improve the antimicrobial efficacy of berberine. Using this approach, six flavonoids not previously identified using synergy-directed fractionation approaches alone (14) were identified as putative additives or synergists. Of these, one compound, predicted by selectivity ratio models to be the top contributor to activity, was isolated and characterized for the first time and its activity as a synergist confirmed. Notably, this compound possessed no antimicrobial activity on its own and may have been missed using biochemometric analyses guided by antimicrobial data alone (18).

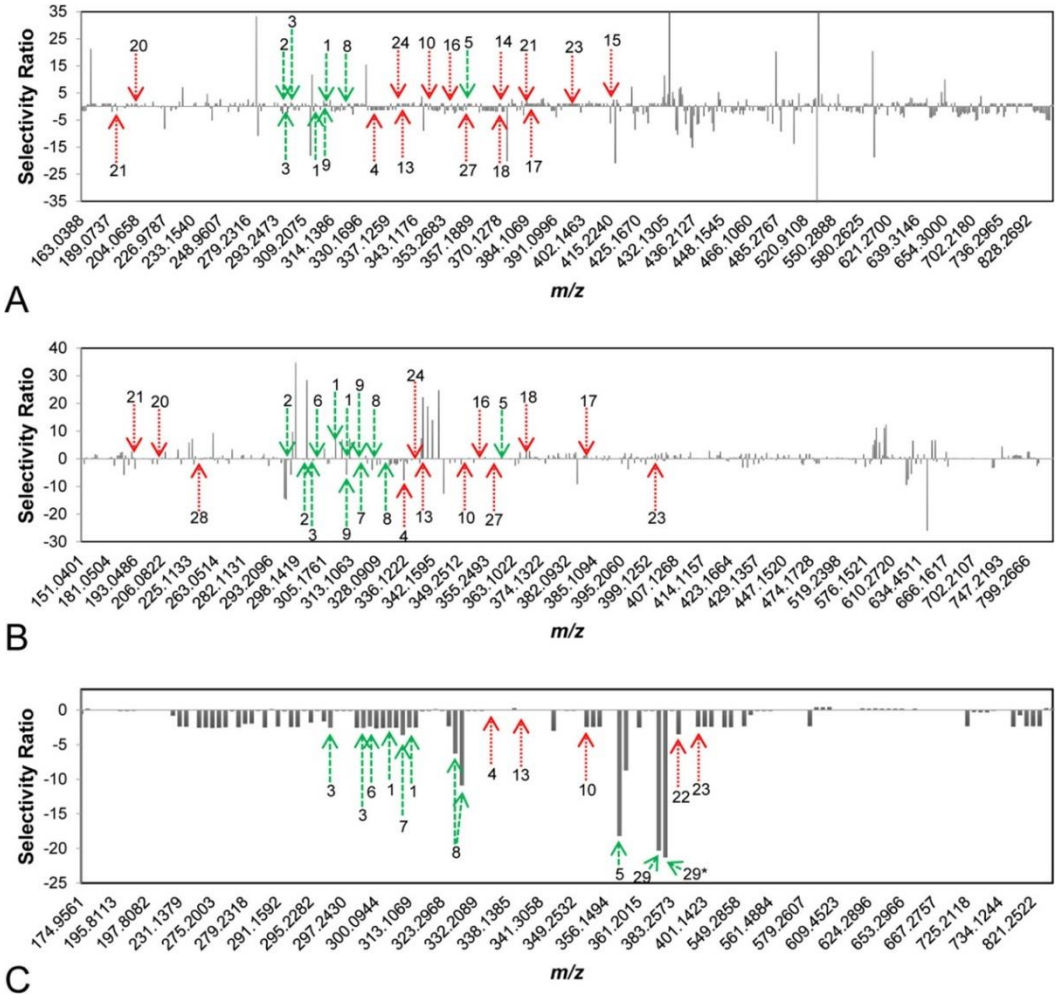


Figure 8. Selectivity Ratio plots for First, Second, and Third Stages of Fractionation [(A), (B), and (C), Respectively] of the Botanical *Hydrastis canadensis.* Growth inhibition data were used to guide selectivity ratio analysis, so variables with negative selectivity ratio are most likely to possess additive or synergistic activity. Known flavonoids (likely to be synergists) are marked in green, while known alkaloids (likely to be additives) are marked in red. First-stage (A) and second-stage (B) models were not able to identify known compounds as contributing to activity. However, the third-stage model (C) predicted seven flavonoids (**1**, **2**, **3**, **5**, **6**, **8**, **29**) and three alkaloids (**10**, **22**, **23**) to possess additive or synergistic activity. With this approach, a new synergistic flavonoid (**29**) was identified in *H. canadensis*, and known flavonoids and alkaloids not previously known to possess additive or synergistic activity were prioritized for future studies. This figure is reprinted with permission from Britton et al. 2017 (18).

Elucidating Mechanisms that Underlie Synergy and Antagonism

In addition to identifying putative active constituents contributing to biological effects of complex mixtures and recognizing the type of interactions they are involved in, it is important to understand the cellular and molecular mechanisms by which complex mixtures exert their effects. To ascertain the molecular targets of mixtures, direct and indirect approaches can be taken.(151) The direct approach utilizes targeted biological assays to identify molecules that affect specific molecular targets while indirect approaches aim to identify mechanisms of action through the evaluation of changes in gene, protein, and/or metabolite profiles in an untargeted manner.(151) While these technologies show great promise, their effectiveness for identifying mechanisms of synergy and antagonism remains to determined.

Targeted assays evaluating specific mechanisms of action (direct approaches)

Targeted approaches to identify mechanisms of action rely on appropriate *in vitro* and *in vivo* models. One important example involves identifying compounds that synergize with existing antibiotics through the inhibition of bacterial efflux pumps (15, 127, 152). A popular method for evaluating efflux pump inhibition involves the use of an efflux pump substrate (such as ethidium bromide or Nile Red) that fluoresces upon contact with cellular DNA (152-154). When efflux pumps are inhibited, fluorescence of the substrate increases due to increased cellular accumulation. This approach has been successfully utilized in numerous studies to identify efflux pump inhibitors from complex botanical mixtures (15, 127). While often successful, these fluorescence-based methods are subject to false results due to matrix quenching effects, particularly when screening

complex natural product mixtures (152). Fluorescence quenching is so common in the biological evaluation of drug candidates that fluorescence quenchers have been tagged as one type of "PAINS" (pan-assay interference compounds) (155, 156). However, the ability to absorb UV/Vis light (and quench fluorescence) is a common feature of druggable small molecules (for example, tetracycline antibiotics) and only constitutes a problem with fluorescence assays. To overcome this limitation, mass spectrometric assays have been developed to monitor efflux pump inhibition or cellular accumulation in *Staphylococcus aureus* (152), *Escherichia coli* (157, 158), *Bacillus subtilis* (157), and *Mycobacterium smegmatis* (157). These assays also offer the distinct advantage of being able to monitor drug accumulation of molecules that do not fluoresce.

Efflux pump inhibition assays, like many other assays used in classical drug discovery approaches, test compounds or mixtures one at a time to identify compounds with promising biological activity. To improve efficiency of these methods, mixtures of drugs can be simultaneously evaluated, but identifying which molecules in these mixtures exert biological effects can be challenging (159). To overcome this limitation, Pulsed ultrafiltration mass spectrometry (PUF-MS) was developed, which enables screening of mixtures such as natural products and synthetic combinatorial libraries (159). PUF-MS involves the incubation of small molecule mixtures with a target protein in solution. Those molecules with affinity for the target will bind to the protein, and compounds that are not bound can be washed away using an ultrafiltration membrane (159). This approach, though effective, is slowed by the ultrafiltration step. To improve the speed of screening, a Magnetic Microbead Affinity Selection Screening (MagMASS) protocol was

developed, in which the protein target of interest is not free in solution, but rather is bound to magnetic beads (160). To separate compounds with and without affinity for the given target, the receptor-bound fraction can be held in solution using a magnet (160). In a recent study, PUF-MS and MagMASS were compared, and both screening methods were found to reliably identify ligands of a specific molecular target from complex botanical matrices (160). MagMASS showed a 6-fold faster separation of bound and unbound compounds when compared to PUF-MS and is compatible with a 96-well plate format (160). Notably, these methods do not require molecules to bind to a particular active site on the target of interest, and can identify ligands that bind to active or allosteric sites. In this way, the assay could be modified to identify combination effects in which the protein's activity is changed through allosteric activation or inhibition (160). However, given that this approach utilizes protein targets rather than whole cells, the combination effects discovered may not translate to intact biological systems. Furthermore, these approaches require access to purified material of the protein target of interest. Therefore, methods based on PUF-MS and its iterations are not applicable for situations where the target of the active compound is either not known or not available.

Indirect approaches to identify multiple targets

While targeted approaches may be useful for identifying compounds that act upon specific molecular targets, assays involving single targets only are not capable of identifying combination effects that involve multiple targets. To identify these multi-target effects, whether it be for a single compound or a combination of multiple constituents, indirect approaches are particularly useful. As discussed in a recent review,

synergistic drug combinations and their modes of action have been explored using molecular interaction profiles (67), and investigation of herbal ingredients using molecular interaction profiles may enable detection of synergistic mechanisms of action. At the time of the review, over 1800 active ingredients from more than 1200 herbs had been subjected to molecular interaction profiling and found to interact with nearly 1000 proteins, many of which were therapeutic targets (67). Although these connections can be utilized to detect potential synergies, the efficacy of complex natural product mixtures and their impact on molecular targets can be influenced by variations in genetics, environment, host behaviour, and timing and dosage of treatment (67). These tools should, therefore, be considered hypothesis-generating, providing a framework for more comprehensive assessment.

The use of DNA and RNA microarrays is another popular tool for probing combination effects within complex mixtures, enabling identification of genes that are up- or down-regulated by natural product extracts alone and in combination. In a recent study, an RNA microarray of neuroglia cells was utilized to compare the number of genes impacted by treatment with *Andrographis paniculata, Eleutherococcus senticosus,* and their fixed combination Kan Jang (161). Results illustrated that *A. paniculata* and *E. senticosus* deregulated 211 and 207 genes, respectively, 36 of which were common to cells treated with each extract (Figure 9A). Using this information, researchers expected that 382 genes would be deregulated in cells treated with the fixed combination Kan Jang. However, only 250 genes were deregulated in Kan Jang treated cells, 111 of which were unique to the Kan Jang combination, potentially due to synergistic interactions

between *A. paniculata* and *E. senticosus.* Alternatively, 170 genes were only affected by treatments with *A. paniculata* or *E. senticosus* and not by the Kan Jang mixture, possibly due to antagonistic interactions between the plant species when applied in combination (Figure 9B) (161). Importantly, microarray analyses do not provide infallible evidence that genes induced by treatments are responsible for physiological effects or mechanisms of synergy, but provide a framework for future research.

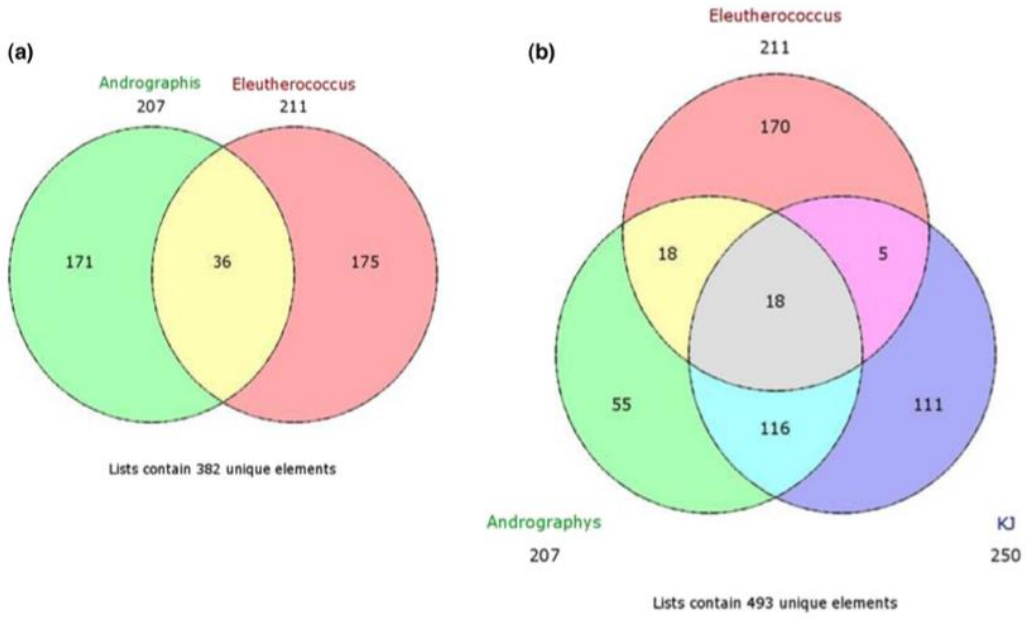


Figure 9. Venn Diagrams of Genes Induced by Treatment of Neuroglia Cells with (a) *Andrographis paniculata* and *Eleutherococcus senticosus* Alone, and (b) *A. paniculata* and *E. senticosus* Alone and in Combination (Kan Jang, Abbreviation KJ). 111 genes are unique deregulated by the Kan Jang mixture, indicating potential gene targets affected by synergistic interactions. The 55 and 170 genes deregulated only by *A. paniculata* and *E. senticosus,* respectively, represent genes potentially impacted by antagonistic interactions in the Kan Jang combination. Reprinted with permission from Panossian et al. 2015 (161).

Because of the material- and time-consuming nature of biological testing, *in silico* approaches have been developed that enable prediction of activity and mechanism of action without the need for direct biological testing. Existing experimental activity data

can be used to mine ligand-target relationships and reveal potential biological activities of diverse molecules (162). Key to the success of this approach for identifying putative mechanisms of action is the availability of compound databases that will facilitate sharing of data and innovation in drug discovery research with both single-target and multi-target approaches (162). Similarly, computational approaches including molecular docking, pharmacophore modelling, and similarity searching can be used as so-called "virtual screening" techniques to identify candidate compounds for follow-up testing (162). Of course, these techniques are subject to error and may not provide accurate representation of the biological system in question, particularly if the model datasets are based on incorrect literature-based annotations of compound activities and/or incomplete understanding of molecular processes of disease.

A systems biology-based approach, network pharmacology, predicts the complex interactions between small molecules and proteins in a biological system, and shows potential as a way to evaluate pharmacological effects of natural product mixtures (162). Unlike the classic "silver bullet" approach where single-target mechanisms are identified for single drugs, network pharmacology focuses on multiple constituents with multiple targets. Several studies have successfully utilized network pharmacology to putatively identify both known and unknown molecular targets (151, 162). Networks can be built using existing literature data, computationally-derived data, or experimental data. The predictive accuracy of the resulting networks relies on the completeness of databases, the robustness of the computational models, the understanding of the underlying mechanisms of disease, and/or the chosen biological assay (162).

Recently, a broad-scale approach was developed in which Functional Signature Ontology (FUSION) maps are utilized to classify putative mechanisms of action of natural products (163). With this method, cellular responses to natural product treatment can be tracked by measuring gene expression of a small, representative subset of genes that provide insight into the physiological state of the cell. The resulting data can then be combined into FUSION maps capable of linking putative bioactive molecules to the proteins and biological pathways that they target in cells (163). This approach has been successfully utilized to link natural products to their mechanisms of action (163) and to identify a marine-derived natural product that inhibits AMPK kinase activity in colon tumor cells (164).

A similar approach, the Connectivity Map, or CMap, was developed in which genes, drugs, and disease states are connected based on the gene expression fingerprints that they share (165). Originally produced using 164 drugs and mRNA expression profiling, the CMap has since been expanded more than 1,000-fold, and now contains over 1.3 million publicly available profiles. This scale-up was achieved using a high-throughput, reduced representation in which only 1,000 landmarks are assessed rather than the full transcriptome. This approach, termed L1000, is sufficient to recover 81% of the information contained in the full transcriptome. The L1000 approach offers advantages over popular approaches such as gene expression microarrays and RNA sequencing because of its low cost and hybridization-based nature, making detection of low-abundant transcripts possible without the need for deep sequencing (165). Preliminary testing has illustrated the potential of the expanded L1000 CMap to

determine the mechanisms of action of small molecules based on the similarities of their genetic perturbations to those of compounds with known activities. This approach can also be utilized to identify potential off-target effects of a drug or drug combinations (165).

During a pilot study, the L1000 CMap was successfully utilized to recover known mechanisms of action from 63% of existing drugs under analysis, to identify the mechanism of action of a previously uncharacterized compound, and to identify compounds with a particular activity of interest. Importantly, the L1000 CMap is not infallible, and 37% of small molecules with known mechanisms of action were not linked to their expected targets during this study. The authors suggest six reasons for this failure: (1) incomplete inhibition of the target by the compound, (2) off target effects, (3) incomplete information in the L1000 data, (4) incorrect data in the literature, (5) biological differences between complete loss of function and loss of a specific protein function, and (6) existence of previously unrecognized connections with stronger connections than expected ones (165). Despite these limitations, the preliminary results of this study emphasize the potential of the L1000 CMap as a launching point for both target-and ligand-based drug discovery (165). Although they have not been explicitly applied to identify mechanisms of synergy or antagonism, the utilization of FUSION maps and the L1000 CMap platform may represent useful tools to enable identification of genes and pathways impacted by a synergistic/antagonistic combination, providing insight into potential mechanisms of action in complex natural product mixtures.

Conclusions and Future Directions

In recent years, the concept of synergy in natural product mixtures has gained attention, and the importance of multi-target combination therapies has come to the forefront. However, the classification of combination effects within complex mixtures and the identification of contributing constituents remains a challenging task, particularly when the majority of established tools have been designed to reduce complexity of natural product mixtures. Additionally, there remains a lack of consensus in the field about which reference models are best for defining combination effects, making interpretation of studies challenging. Recent models using the Explicit Mean Equation (46) and the Zero Interaction Potency model (38) represent newly developed and robust reference models that may permit improved identification combination effects. These models have yet to be employed for real world applications in studying natural product mixtures, and future studies will reveal their applicability for this approach.

Metabolomics and biochemometric approaches are promising tools for studying synergy, and have just begun to be applied to identifying constituents that participate in combination effects.(18) While useful, biochemometric models are subject to limitations based on the biological assays and reference models used to define biological activity. Similarly, the linear regression models used to predict active constituents are inherently limited given that true linear relationships rarely exist, particularly when assessing complex mixtures with numerous unknown combination effects. The application of statistical tools capable of identifying non-linear relationships will be helpful for future research initiatives. In addition, untargeted approaches to identify molecular targets of

synergy and unravel synergistic (or antagonistic) mechanisms of action have just begun to be explored, and continued studies on this topic are of the utmost importance. Advancements in Big Data approaches show great promise for identifying active mixture constituents, characterizing the nature of their interactions, and elucidating their potential mechanisms of action. Integrated technologies capable of completing all of these tasks simultaneously remain to be developed. The production of such integrated techniques will become increasingly important in our continued pursuit to understand the biological activities of complex mixtures.

Acknowledgements

This research was supported by the National Center for Complementary and Integrative Health of the National Institutes of Health under grant numbers 5 T32 AT008938 and 1 R15 AT010191. Ashley Scott is also acknowledged for her help preparing figures for this manuscript.

CHAPTER II

A REVIEW OF THE MEDICINAL USES AND PHARMACOLOGY OF ASHITABA

This chapter has been published in the journal Planta Medica and is presented in that style. Caesar, L.K., Cech, N.B. Planta Medica. 2016, 82(14), 1236-1245.

Caesar, L.K. searched all of the available literature and wrote the manuscript. Cech, N.B. provided suggestions and edits throughout manuscript preparation.

Introduction

Complementary health practices are gaining global popularity, and a recent National Health Interview Survey estimated that nearly 18% of adults in the United States regularly took non-vitamin, non-mineral dietary supplements in 2012 (7). Because of the growing popularity of herbal medicine, it is important to understand the chemical basis behind the purported activities of botanicals. *Angelica,* a member of the Apiaceae (Umbelliferae) family, is a large genus comprised of over 60 species. Members of the genus have been utilized as medicines across the world, most notably in Asia, to treat numerous ailments, including influenza, hepatitis, arthritis, indigestion, fever, and microbial infections (166). An increasing number of studies are being conducted on a medicinally promising member of the genus, *Angelica keiskei* Koidzumi (Apiaceae), or ashitaba. This large leafy perennial plant native to the Pacific coast of Japan is used throughout Asia for its diuretic, laxative, stimulant, and galactagogue properties (167). In

the past decade, several active constituents representing chalcones, flavanones, and coumarins, have been isolated and characterized from ashitaba, and several bioactivities have been described. This review presents the current progress on ashitaba pharmacological studies, with focus on isolated secondary metabolites, biological activity, toxicological data, and clinical relevance.

Bioactive Metabolites Isolated from Ashitaba

Chalcones

Most of the literature on bioactive metabolites from ashitaba concerns the diverse activity of various chalcones (Table 2; Figure 10), which are most abundant in the root bark of the plant (168). Chalcones are formed from phenylpropanoid starter units, extended with three malonyl-CoA molecules. The resulting polyketide is folded by the enzyme chalcone synthase to promote Claisen condensations and subsequent enolizations (169). Interestingly, the bioactive chalcones found in ashitaba are prenylated at the 5'-position (Figure 10), indicating that these molecules have undergone multiple biosynthetic steps, travelling through the acetate, shikimate, and isoprenoid pathways.

Many chalcones, both from ashitaba and other natural product sources, have been shown to possess chemopreventive, anti-diabetic, antibacterial, anti-inflammatory, and anxiolytic properties, as well as others (170-174). In many instances, a single chalcone may demonstrate multiple bioactive properties. These diverse bioactivities may be attributed to the flexible structural conformation of the chalcone backbone, leading to promiscuous substrate behavior (175). Two chalcones, 4-hydroxyderricin (**1**) and

xanthoangelol (**2**) are most abundant in this plant, and possess cytotoxic, anti-inflammatory, and anti-diabetic properties (176).

Table 2. Isolated Bioactive Components from *Angelica keiskei* Koidzumi and Plant Part from which they were First Isolated.

No.	Compound name	Part of plant	References
Chalcones			
1	4-hydroxyderricin	Roots	(177)
2	xanthoangelol	Roots	(177)
3	xanthoangelol B	Roots	(178)
4	xanthoangelol C	Roots	(178)
5	xanthoangelol D	Roots	(178)
6	xanthoangelol E	Roots	(178)
7	xanthoangelol F	Roots	(179)
8	xanthoangelol G	Roots	(179)
9	xanthoangelol H	Roots	(179)
10	xanthoangelol I	Stems	(167)
11	xanthoangelol J	Stems	(167)
12	xanthoangelol K	Stems	(180)
13	xanthokeistal A	Leaves [a]	(181)
14	isobavachalcone	Roots	(179)
15	(2E)-1-[3,5-Dihydroxy-2-methyl-2-(4-methyl-3-penten-1-yl)-3,4-dihydro-2H-chromen-8-yl]-3-(4-hydroxyphenyl)-2-propen-1-one	Roots	(182)
16	(2E)-1-[4-Hydroxy-2-(2-hydroxy-2-propanyl)-2,3-dihydro-1-benzofuran-7-yl]-3-(4-hydroxyphenyl)-2-propen-1-one	Roots	(182)
17	(2E)-1-[4-Hydroxy-2-(2-hydroxy-6-methyl-5-hepten-2-yl)-2,3-dihydro-1-benzofuran-5-yl]-3-(4-hydroxyphenyl)-2-propen-1-one	Roots	(182)
18	(2E)-1-(3-[(2E)-6,7-Dihydroxy-3,7-dimethyl-2-octen-1-yl]-2,4-dihydroxyphenyl)-3-(4-hydroxyphenyl)-2-propen-1-one	Roots	(182)
19	(2E)-1-(3-[(2E)-6-Hydroperoxy-3,7-dimethyl-2,7-octadien-1-yl]-2-hydroxy-4-methoxyphenyl)-3-(4-hydroxyphenyl)-2-propen-1-one	Roots	(182) (183)
20	xanthokeismin A	Stems	(183)
21	xanthokeismin B	Stems	(183)
22	xanthokeismin C	Stems	
Coumarins			
23	(3'R)-3'-hydroxy-columbianidin	Stems	(184)
24	3'-senecioyl khellactone	Stems	(184)
25	5-methoxypsoralen	Fruit	(178)
26	4'-senecioyl khellactone	Stems	
27	archangelicin	Fruit	(178)
28	isolaserpitin [b]	Fruit	(178)
29	laserpitin [b]	Fruit	(178)
30	osthenol	Stems	(167)

31	pteryxin	Stems	(184)
32	demethylsuberosin	Aerial portion	(185)
33	selinidin	Fruit	(178)
Flavanones			
34	8-geranylnaringenin	Stems	(167)
35	4'-O-geranylnaringenin	Stems	(184)
36	Isobavachin	Stems	(167)
37	munduleaflavanone	Stems	(184)
38	munduleaflavanone B	Stems	(167)
39	prostratol F	Stems	(184)
Other compounds			
40	ashitabaol A	Seeds	(186)
41	falcarindiol	Stems	(184)
42	pregnenolone	Aerial portion	(185)
43	4-hydroxy-3,5,5-trimethyl-4-(1,2,3,-trihydroxybutyl)cyclohex-2-enone	Aerial portion	(185)

[a] part of plant was inferred, but not directly stated by authors.
[b] common names laserpitin and isolaserpitin also refer to sesquiterpene-type compounds. In this case, they refer to angular coumarin derivatives isolated from Ashitaba fruits. Other references cited in this review utilize this nomenclature, as well.

Coumarins

Ashitaba contains numerous coumarins with medicinal properties (Table 2; Figure 11). Coumarins result from the addition of an hydroxy- group *ortho*- or *para*- to the propanoid side chain of cinnamic acids (187). Although basic coumarins are comprised solely of a phenyl-propanoid backbone with varying degrees of hydroxylation, many others have more complex carbon frameworks derived from isoprene units. These 5-carbon units can lead to cyclization with a phenol group, eventually yielding complex coumarin derivatives (187). Depending on the position of the initial dimethylallylation, furocoumarin derivatives may be angular (**23**, **24**, **26-29**, **31**, **33**), or linear (**25**).

Coumarins isolated from a number of plant species have been shown to possess anti-inflammatory and chemopreventive properties (188, 189). Indeed, coumarins isolated from ashitaba have demonstrated cytotoxic properties (167, 184, 190), in

addition to anti-diabetic (180), anti-obesity (176), and blood pressure reducing effects (191).

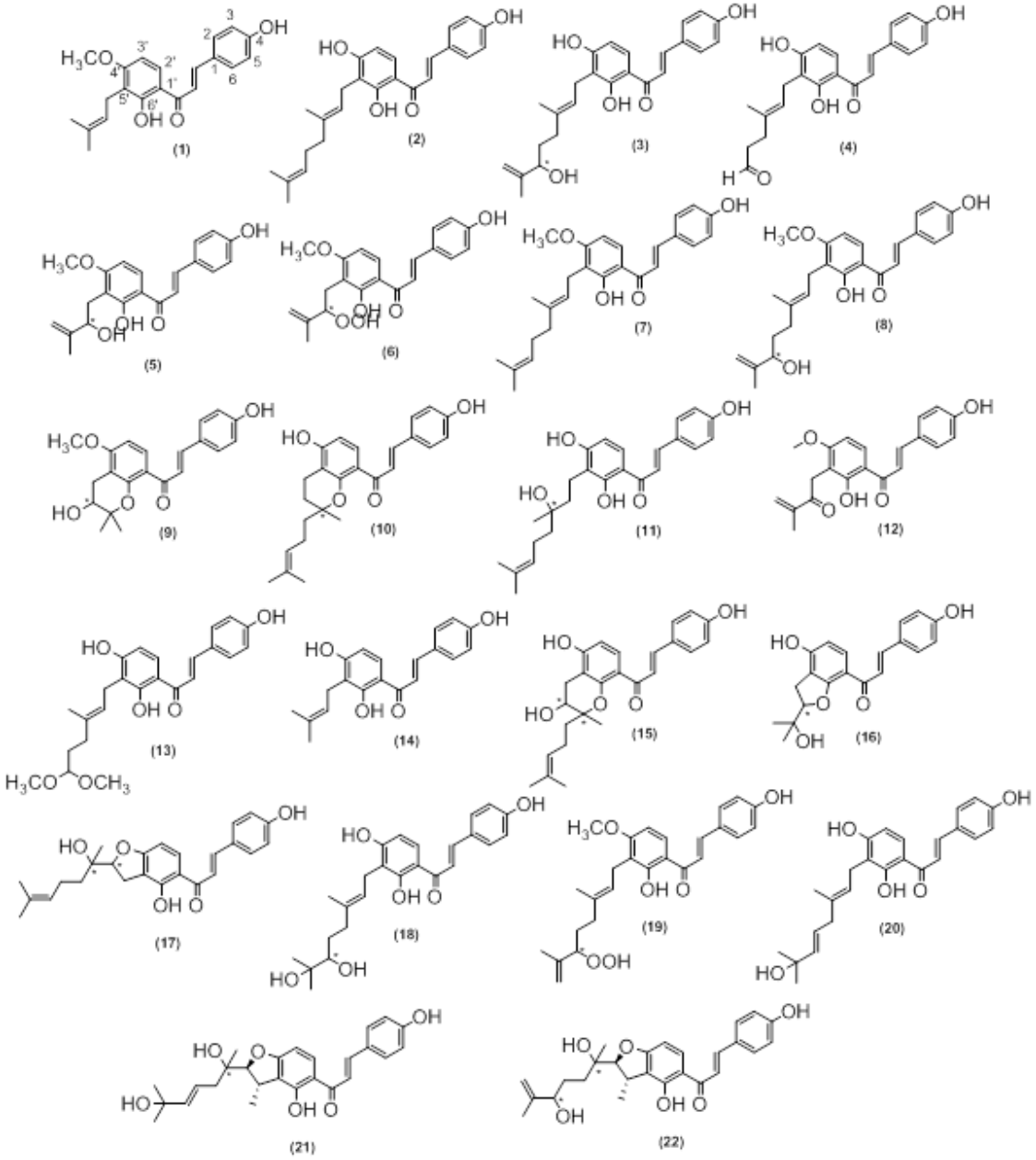


Figure 10. Structures of Chalcones Isolated from *Angelica keiskei* Koidzumi. Absolute configuration at points marked with an asterisk (*) were not specified in original articles.

Flavanones

Considering the abundance of chalcones found in ashitaba, it is not surprising that this plant also possesses several flavanones (Table 2; Figure 12). Chalcones, with a nucleophilic phenol group positioned near to an α,β-unsaturated ketone readily undergo Michael-type attack, leading to cyclization and flavanone formation (192).

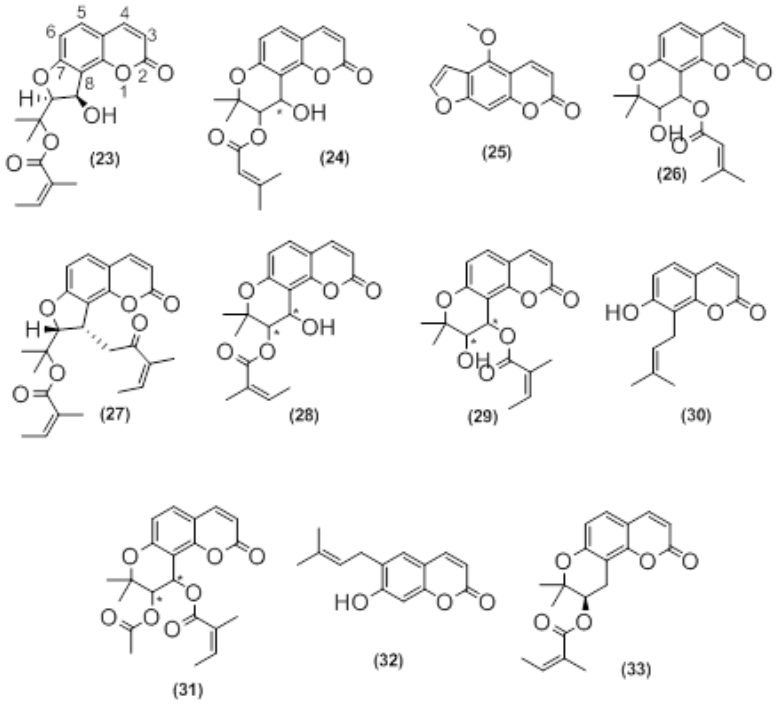


Figure 11. Structures of Coumarins Isolated from *Angelica keiskei* Koidzumi. Absolute configuration at points marked with an asterisk (*) were not specified in original articles.

Flavanones are distributed throughout the plant kingdom and are found in 42 plant families both in aerial and below ground tissue. These compounds have been shown to possess radical-scavenging, anti-inflammatory, and chemopreventive effects (193). Flavanones in ashitaba, though less studied than the chalcones **1** and **2**, have been studied most for their potential as chemopreventive agents (184).

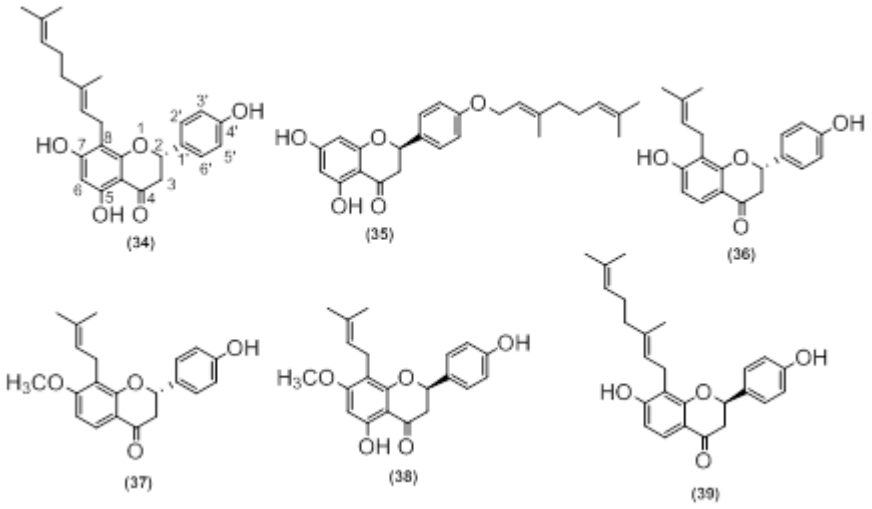


Figure 12. Structures of Flavanones Isolated from *Angelica keiskei* Koidzumi.

Other active compounds

Ashitaba also possesses active polyacetylenes, triterpenes, and cyclohexenones. One sesquiterpene, ashitabaol A (**40**) has been isolated from ashitaba seeds (Table 2, Figure 13) and shows free radical scavenging activity (186). Sesquiterpenes containing a hexahydrobenzofuran or tetrahydro-backbone with the 3-methyl-but-2-enylidene unit are extremely uncommon in nature. Compound **40** is only the second reported natural product, after bisbolangelone, with this unusual structure (186).

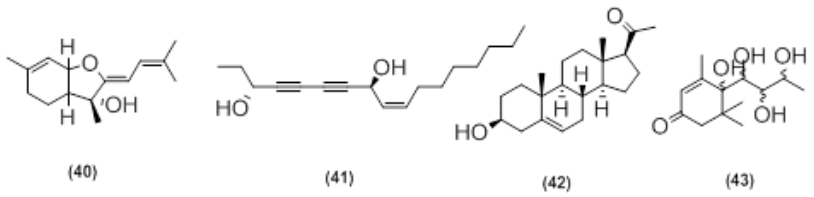


Figure 13. Other Compounds Isolated from *Angelica keiskei* Koidzumi.

Biological Activities of Ashitaba

Extracts of ashitaba, whether containing complex mixtures or isolated compounds, are used to treat many diseases. In this section we describe ashitaba's cytotoxic, anti-diabetic, anti-obesity, antioxidant, anti-inflammatory, antithrombotic, antihypertensive, and antimicrobial properties. When possible, structure-activity relationships of known active constituents will be described. A summary of the *in vivo* and *in vitro* studies on ashitaba extracts can be found in Table 3. A comprehensive list of known bioactivities for each isolated compound can be found in Table 4.

Table 3. *In vitro* and *in vivo* Bioactivity Studies on Ashitaba Extracts.

Plant Part	Extract Type	Biological Activity Tested	Results	References
Cytotoxicity				
Not specified[a]	Ethyl acetate extract	Anticarcinogenicity (*in vitro)*	Hep G2 cells treated with ashitaba extract (1 mg mL^{-1}) showed a 1.42-fold induction of quinone reductase expression, an anticarcinogenic marker enzyme.	(194)
Fresh aerial portion	95% ethanol extract	Anticarcinogenicity (*in vitro*)	Murine hepatoma Hepa 1c1c7 cells treated with 25 µg mL^{-1} ashitaba extract showed a 2.44-fold induction of NAD(P)H quinone oxidoreductase 1, protecting against quinone-induced damage.	(185)
Anti-diabetic and Anti-obesity activity				
Stem exudate	Ethyl acetate extract	Anti-hyperlipidemic (*in vivo)*	Male stroke-prone spontaneously hypertensive rats fed a diet containing 0.2% ashitaba extract for 6 weeks showed increased levels of serum HDL levels and reduced liver triglyceride levels correlated with down-regulation of hepatic acyl-coenzyme A synthetase mRNA.	(195)

Leaves and processed products of leaves	Whole leaves, juice, fermented juice, and/or squeeze debris	Anti-adiposity (*in vivo*)	Male Sprague-Dawley rats fed a high fat diet with 3-5% ashitaba whole leaves or a combination of juice and solid squeeze debris for 6 weeks showed decreased liver, kidney, epididymal fat, and rear fat weights. Ashitaba and its processed products increased luteolin absorption and suppressed diet-induced cholesterol build up in the liver by increasing antioxidant enzyme gene expression.	(196)
Stem exudate	Ethyl acetate extract	Anti-adiposity (*in vivo)*	Male C57BL/6 mice fed a high-fat diet with 0.01% ashitaba extract by weight for 16 weeks showed lowered diet-induced body weight and body fat and lowered serum levels of glucose, insulin, and cholesterol when compared to positive controls. Ashitaba extract regulated lipid metabolism in adipose and liver tissue by activating AMP-activated protein kinase.	(176)
Not specified[b]	Ashitaba powder	Anti-adiposity (*in vivo)*	Male Wistar rats fed a high fat diet in combination with ashitaba powder at 17, 170, or 1700 mg 100 g^{-1} body weight for 28 days did not show significant differences in body weight gain, food intake, or relative organ weights when compared to positive controls.	(197)
Dried leaves and stems	Ethanol extract	Anti-diabetic (*in vivo*)	Male Wistar rats fed a high-fructose diet with 3% ashitaba extract by weight for 11 weeks had 16.5% lower blood glucose levels, 47.3% lower serum insulin, 56.4% lower HOMA-R, and 24.2% lower triglyceride content, leading to improved insulin resistance and hypertriglyceridemia when compared to positive controls, likely by enhancing expression of genes related to β-oxidation of fatty acids. .	(198)
Roots	Ethanol extract	Anti-diabetic (*in vitro*)	Ashitaba extract showed insulin-like activity following incubation with 3T3-L1 cells. Dose-dependent glucose uptake and differentiation of preadipocytes to adipocytes were observed in treated cells but not in controls.	(199)
Anti-inflammatory activity				
root cores,	Methanol extract	Xanthine oxidase inhibition	Xanthine oxidase enzyme from bovine serum milk inoculated with 3.12, 6.25,	(168)

root bark, leaves, and stems		(*in vitro*)	and 12.5 µM of 4 extracts and 20 mM xanthine was assayed by tracking xanthine oxidation spectrophotometrically. Extracts all showed lower OD_{273} values than the positive control, allopurinol, indicating that all extracts had potent XO inhibitory activity. Stem and root bark extracts were the most potent inhibitors.	
Not specified	n-hexane extract	Anti-inflammatory (*in vitro*)	Ashitaba extract 10, 30, 50, or 100 µg mL^{-1} suppressed lipopolysaccharide-induced JNK, p38, and ERK1/2 activation in RAW 264.7 macrophages. NF-KB was suppressed as well through inhibition of p65 translocation and phosphorylation.	(200)
Stem exudate	Yellow exudate, ethyl acetate extract, chalcone-rich, and coumarin-rich fractions	Anti-inflammatory (*in vivo*)	Male kwl ICR mice (pathogen free grade) injected intraperitoneally with Ashitaba exudate for 7 days before injection with lipopolysaccharide significantly inhibited increase of PAI-1 antigen in lung and liver tissue at 6 and 9 hours. Additionally, ethyl acetate extract and chalcone-rich fractions decreased production of LPS-induced PAI-1.	(201)
Antihypertensive activity				
Freeze dried leaves	Purified fraction from 80% ethanol crude extract	Antihypertensive (*in vivo*)	Male spontaneously hypertensive rats given ashitaba extract at 21.8 mg kg^{-1} a day for 10 weeks showed significantly lower blood pressure (200 ± 7.3 mmHg) when compared to control rats (211 ± 3.7 mmHg)	(202)

[a] edible parts of washed vegetables
[b] "Ashitaba powder commercially available as a so-called functional food"

Table 4. Bioactivities Attributed to Compounds Isolated from Ashitaba.

Compound	Bioactivities	References
1	Chemopreventive, antidiabetic, anti-adipogenic, anti-inflammatory, anti-platelet, anti-influenza, antibacterial	(168, 176, 180-182, 184, 185, 199, 203-209) (210)
2	Chemopreventive, antidiabetic, anti-adipogenic, anti-inflammatory, antioxidant, anti-platelet, antibacterial	(168, 176, 180, 182-185, 199, 201, 203, 205, 208, 209) (204, 206, 210)

3	Anti-inflammatory, antioxidant, anti-platelet, anti-influenza	(168, 181, 183, 201, 210)
4	Anti-inflammatory	(210)
5	Anti-diabetic, anti-inflammatory, anti-influenza	(180, 181, 201, 207)
6	Anti-diabetic, anti-inflammatory, anti-platelet,	(180), (210) (201)
7	Chemopreventive, anti-diabetic, anti-inflammatory, antioxidant; anti-influenza	(167, 168, 180, 181, 184, 203)
8	Anti-influenza	(181)
9	Chemopreventive	(184)
10	Chemopreventive, anti-inflammatory	(167)
11	Chemopreventive, anti-inflammatory	(167)
12	Anti-diabetic	(180)
13	Anti-influenza	(181)
14	Chemopreventive, anti-inflammatory	(167, 168, 184, 185, 203)
15	Anti-diabetic, antioxidant	(182, 185)
16	Anti-diabetic	(182)
17	Anti-diabetic	(182, 185)
18	Chemopreventive, anti-diabetic	(182)
19	Anti-diabetic	(182)
20	Antioxidant	(183)
21	Antioxidant	(183)
22	Antioxidant	(183)
23	Chemopreventive	(184)
24	Chemopreventive; anti-inflammatory	(167, 184)
25	Anti-diabetic	(180)
26	Chemopreventive, anti-inflammatory	(167, 184)
27	Chemopreventive	(190)

28	Chemopreventive, anti-inflammatory	(167, 184)
29	Chemopreventive, anti-inflammatory	(167, 184)
30	Chemopreventive, anti-inflammatory	(167, 184)
31	Chemopreventive, anti-inflammatory	(167, 184)
32	Anti-diabetic	(185)
33	Anti-inflammatory	(167, 184)
34	Chemopreventive, anti-inflammatory	(167, 184)
35	Chemopreventive	(184)
36	Chemopreventive	(167)
37	Chemopreventive	(184)
38	Chemopreventive, anti-inflammatory	(167, 184)
39	Chemopreventive	(184)
40	Antioxidant	(199)
41	Anti-diabetic	(185)
42	Anti-oxidant	(183)
43	Anti-oxidant	(183)

Antidiabetic and anti-obesity activities

Although ashitaba has been purported to possess numerous bioactivities, it has most notably been utilized as a medicinal plant to prevent obesity and its complications. Ashitaba extracts and their isolated constituents have been shown to possess antidiabetic and anti-obesity properties. However, the purported properties and modes of action are often contradictory between studies, suggesting a need for more comprehensive analysis of these activities.

Tyrosine-protein phosphatase 1B (PTP1B) negatively regulates the insulin signaling pathway, and is a promising target for the treatment of type-II diabetes mellitus (180). Several compounds isolated from ashitaba, including chalcones **1**, **2**, **5**, **6**, **7**, and **12** and a coumarin (**25**), inhibited PTP1B activity with IC_{50} values of 0.82-4.42 μg mL^{-1}. Kinetic studies revealed that compound **12** was a fast-binding competitive inhibitor of PTP1B (180). Additionally, KK-A^y mice, known to develop hyperglycemia with aging, were fed diets comprised of 0.15% **1** or **2** and showed suppressed development of insulin resistance, as well as lower levels of blood glucose (50% and 33% lower, respectively) when compared to controls (199).

Alpha-glucosidases aid in carbohydrate digestion and glucose release, and increased activity of these enzymes can lead to hyperglycemia and the development of type-II diabetes. Alpha glucosidase inhibitors are target molecules for suppressing the onset of this disorder. Four compounds, **14**, **32**, and **41**, had alpha glucosidase inhibitory activity with IC_{50} values below 20 μM when using 4-nitrophenyl-alpha-D-glucopyranoside as the substrate, considerably lower than the control drug acarbose (IC_{50} = 384 μM) (185).

To maintain blood sugar homeostasis, it is imperative that skeletal muscle cells uptake glucose. Obesity can impair this uptake and lead to hyperglycemia. The majority of the translocation of glucose is completed by glucose transporter 4 (GLUT4). The activity of GLUT4 is regulated by protein kinase ζ/λ (PKC ζ/λ), protein kinase B (Akt), and adenosine monosphosphate activated protein kinase (AMPK). The activities of **1** and **2** on the activation of GLUT4 glucose translocation in rat skeletal muscle L6 cells were

determined and compared to the activity induced by insulin (205). At 30 μM, **1** stimulated glucose uptake into L6 myotubes 2.8-fold, and **2** stimulated the uptake 1.9-fold, as did insulin. At 10 μM, **1**, **2**, and insulin induced GLUT4 translocation equally. Of the compounds screened, the prenylated chalcones had the highest GLUT4 inducing activity. The hydrophobic groups may interact directly with the myotubes and facilitate activation of transporters (205). Interestingly, the authors found that proteins that typically induce GLUT4 activity, notably PKC ζ/λ, Akt, and AMPK, were not activated by **1** and **2**. Thus, **1** and **2** affect other signaling components in the cascade.

The differentiation of adipocytes from pre-adipocytes plays a large role in the development of obesity (209). Peroxisome proliferator-activated receptor γ (PPAR-γ) and CCAAT/enhancer binding proteins (C/EBPs) play important regulatory roles in adipocyte differentiation. Activation of C/EBP-β and C/EBP-δ begins a cascade that increases expression of C/EBP-α, PPAR-γ, and GLUT4 (209). AMPK downregulates C/EBP-α and PPAR-γ expression, and modulates the activity of other factors through the inactivation of acetyl-CoA carboxylase (ACC). Inactivation of ACC by phosphorylation halts the biosynthesis of malonyl-CoA, leading to fatty acid oxidation by carnitine palmitoyltransferase-1A (CPT-1A) (176).

Counterintuitively, ligands that activate PPAR-γ have been developed to treat type-II diabetes mellitus. Small adipocytes can enhance glucose uptake upon insulin stimulation, enabling the reduction of insulin resistance (199). One study determined that incubation of 3T3-L1 cells with compounds **1** and **2** instead of insulin led to equal levels of adipocyte differentiation, but compound **1** resulted in the highest induction of glucose

uptake. In a follow-up experiment, the effects of **1** and **2** on PPAR-γ were evaluated, along with the effects of a known PPAR-γ agonist, pioglitazone. Interestingly, only the known agonist pioglitazone activated PPAR-γ, indicating that compounds **1** and **2** induce glucose uptake by a different mechanism than PPAR- γ activation (199).

Other studies have reported contradictory results, and indicate that ashitaba extracts, and particularly compounds **1** and **2** suppress adipocyte differentiation by inactivating PPAR-γ (176, 209). Treatment of 3T3-L1 cells with **1** and **2** phosphorylated AMPK, leading to its activation, and subsequent downregulation of C/EBP-α, C/EBP-β, PPAR-γ, and GLUT4 expression (209). To determine if adipogenesis was inhibited as a result of AMPK activation, cells were treated with compound C, an AMPK inhibitor, and with compounds **1** and **2**. Compound C reversed the anti-adipogenic effects of the chalcones, further supporting the involvement of **1** and **2** in AMPK activation (209).

Adiponectin helps to improve insulin resistance, so compounds aiding in adiponectin production may be useful in inhibiting the development of metabolic syndrome (182). In one study, the effects of compounds **1**, **2**, and **15-19** were assessed for their effects on adiponectin production in 3T3-L1 adipocytes. All chalcones up-regulated expression of adiponectin mRNA, particularly compounds **17** (7.80-fold induction) and **18** (8.27-fold induction). Compounds **1**, **2**, and **15-19** also significantly enhanced adiponectin production (182).

One clinical study was conducted to determine ashitaba's efficacy for treating metabolic syndrome. For this study, 9 subjects ingested ashitaba juice comprised of dried leaves and stems for 8-weeks (211). Following ingestion, all subjects had significantly

lower visceral fat, body fat, and body weight at the end of the 8th week, and no adverse clinical changes were attributed to ashitaba. However, this study lacked controls and as such provides insufficient evidence for ashitaba's efficacy in treating metabolic syndrome.

Numerous *in vitro* and *in vivo* studies support the use of ashitaba as an anti-obesity and anti-diabetic agent, although clinical trials are needed to confirm the relevance of these compounds in humans. However, contradictions in the literature suggest that further research to understand the mechanisms of action and molecular targets of active constituents should be conducted in addition to clinical tests. Additionally, research on other ashitaba constituents besides compounds **1** and **2** may lead to novel discoveries.

Chemopreventive activity

Ashitaba extracts have been shown to possess chemopreventive properties *in vitro*, involving both antiproliferative and antimutagenic mechanisms. Quinone reductase plays an important role in detoxification by reducing electrophilic quinones. This defends cells against quinone-induced cytotoxic effects and subsequent carcinogenesis(194). An ethyl acetate soluble crude vegetable extract of ashitaba was shown to induce Hep G2 cell quinone reductase activity by nearly fifty percent in 48 hours (1.42 ± 0.06 fold induction)(194). Unfortunately, the part of the plant extracted was not specified, and chemical consituents were not determined(194). Another study determined that NAD(P)H quinone oxidoreductase 1 (NQO1), which also protects against quinone-induced damage, was activated in murine hepatoma Hepa 1c1c7 cells by an ethanol soluble extract of

ashitaba (2.44-fold induction at 25 μg mL^{-1}). Subsequent compound isolation indicated that four chalcones, **1**, **2**, **14**, and **18** had the highest rates of NQO1 induction when tested against murine hepatoma Hepa 1c1c7 cells (185).

Several researchers have studied the inhibitory effects of ashitaba compounds on the induction of the Epstein-Barr virus Early Antigen (EBV-EA) by 12-*O*-tetradecanoyl phorbol 13-acetate (TPA). EBV is associated with numerous diseases, including types of lymphoma and cancer, and the inhibitory effects on its induction are often used to evaluate antitumor-promoting activity in preliminary studies (167). In Raji cells, compounds **10**, **11**, **14**, **30**, **34**, **36**, and **38** were more potent inhibitors than retinoic acid, the reference compound, with IC_{50} values ranging from 215-320 mol ratio 32 $pmol^{-1}$ TPA(167). In a previous study, compounds **1**, **2**, **7**, **9**, **23**, **24**, **26**, **28**, **29**, **31**, **35**, **37**, and **39** also showed potent inhibitor effects, ranging from 92-100% inhibition at 1000 mol ratio, and 51-84% at 500 mol ratio. In Raji cells, inhibitors **1**, **14**, **35**, **37**, and **39** were more potent than the reference compound β-carotene (184). Compound **27** was also found to have TPA-inhibiting properties (190). All of these active compounds have, in addition to the chalcone, coumarin, or flavanone backbone, a prenyl or genanyl group, suggesting that the addition of isoprene units results in an increase in chemopreventive potential (167).

Three prenylated chalcones, **1**, **2**, and **7**, were transformed by the fungal microbe *Aspergillus satoi*, resulting in flavanone, prenyl-chain hydrated, and ring-B-hydroxylated derivatives. Several flavanone and prenyl-chain derivatives, along with compounds **1**, **2**, and **7**, also suppressed EBV-EA induction in Raji cells with IC_{50} values ranging from

211-348 mol ratio 32 $pmol^{-1}$ TPA (203). Interestingly, biotransformation products in which the prenyl or geranyl chain was hydrated had the most potent inhibitor effects, even more than parent compounds. Products that had been cyclized from chalcones to flavanones, on the other hand, showed weakened activity (203).

A prenyl-chain hydrated biotransformation product of **1**, 2",3"-dihydro-4,3"-dihydroxyderricin (**44**, Figure 14), was shown to possess cytotoxic activity (IC_{50} = 2.9 µM) against human leukemia cells (HL60) (203). To determine if this compound played a role in regulating apoptosis, a follow up experiment was conducted. Indeed, HL60 cells treated with 30, 40, or 50 µM of this compound displayed morphological characteristics consistent with apoptosis, including chromatin condensation, nuclei fragmentation, and mitochondrial membrane collapse (203). In two-stage carcinogenesis tests in mouse skin, it was determined **14** and 6",7"-dihydro-7"-hydroxyxanthoangelol F (**45**, Figure 14), a hydrated prenyl-chain biotransformation product of **7**, inhibited the rate and number of skin tumors produced in mice . When topically treated twice a week with 7,12-dimethylbenz[*a*]anthracene (DMBA) and TPA, control mice developed papillomas 100% by 11-weeks. When treated topically with 85 nmol of **45** before application of DMBA and TPA, the incidence was lowered to 27% at 11-weeks and 87% at 20 weeks (203). Similarly, after 10-weeks only 20% of mice given a topical treatment (85 nmol) of **14** before contact with tumor-inducing compounds developed papillomas when compared to 100% of controls. At 20-weeks, 87% of treated mice had developed papillomas (167).

Several *in vitro* studies have been conducted on ashitaba's cytotoxic effects. However, only a few *in vivo* tests have been completed using animal models, and no

clinical trials have been conducted in humans. As such, no conclusive evidence yet exists to confirm the use of ashitaba compounds as anticancer agents. More robust animal studies followed by clinical trials are necessary to support the use of these constituents for cancer treatment.

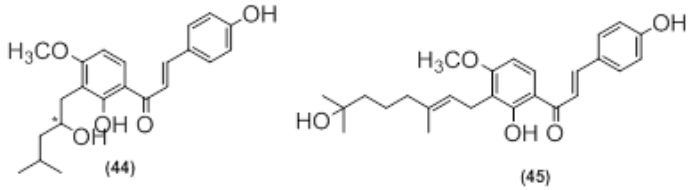


Figure 14. Chalcone Biotransformation Products from *Angelica keiskei* Koidzumi.

Oxidative stress relief and anti-inflammatory activity

Compounds isolated from ashitaba have been shown to possess antioxidant properties, thereby reducing inflammation by a number of routes. Modes of action include xanthine oxidase (XO) inhibition (168), free-radical scavenging activity (183, 185, 186), and reduction in expression of pro-inflammatory transcription factors (200, 207, 208).

Xanthine oxidase (XO) reduces molecular oxygen, leading to anionic O_2^- and hydrogen peroxide. These free radicals commonly result in inflammation, so regulators of XO activity could be potent anti-inflammatory agents (168). When tested against XO from bovine serum milk, ashitaba stem and root bark extract demonstrated significant XO regulation as indicated by increased levels of xanthine oxidation. Isolated chalcones **1**, **2**, **3**, **7**, and **14** showed IC_{50} values against XO ranging from 8.1 to 54.3 µM. Compound **2**

was found to be the most effective (IC_{50} = 8.1 μM) and likely functions as a reversible inhibitor of xanthine oxidase (168).

Generation of free radicals can result in damage to cellular machinery. Compound **40** from ashitaba seed coat tissue exhibited 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulphonic acid (ABTS) free radical scavenging activity (186). Additionally, compounds **2**, **15**, **42**, and **43** were found to scavenge 2,2,-diphenyl-1-picrylhadrazyl (DPPH) radicals (185), indicating that these compounds may be useful antioxidant agents. Compounds **3**, **20**, **21**, and **22** were also shown to scavenge superoxide radicals (0.51-1.1 μM IC_{50} values), with **20** showing the most potent activity (183).

Nitric oxide (NO) is another mutagen that affects microbial and mammalian cells due to the production of free radicals. When tested against Chang liver cells, compounds **7**, **10**, **11**, **14**, **24**, **26**, **28**, **29**-**31**, **33**, **34**, and **38** showed inhibitory effects on NO almost equal to the reference compound glyzyrrhizin (167, 184). In another study, compounds **1** and **2** were also shown to suppress the production of NO in RAW264 macrophages, with negligible effects on cellular function (208). The authors noted that prenylated chalcones were more effective in suppressing NO formation, with **2** being more potent than **1**. Since **2** contains a geranyl group and **1** contains a dimethylallyl group, it is possible that the increased hydrophobicity of additional isoprene units facilitates compound accumulation into the cell, promoting antioxidative activity (208).

Tumor necrosis factor alpha (TNF-α) has been implicated as an important participant in the induction of inflammation (208) and is regulated by transcription factors activator protein 1 (AP-1) and the nuclear factor kappa-light-chain-enhancer of

activated B cells (NF-KB). Ashitaba extract and compound **2** were shown to inhibit inflammation induced by TNF- α in male kwl ICR mice (201). Another study determined that isolated compounds **1** and **2** had similar TNF- α suppressing effects in RAW264 macrophages (208), and compound **5** induced suppression in porcine aortic endothelial cells (207). In RAW246.7 macrophages, the n-hexane ashitaba extract had anti-inflammatory activity resulting from down-regulation of NF- KB-dependent gene products (200). Ashitaba's anti-inflammatory properties can also be attributed to its effects on histamine release. Histamine is an important messenger compound released by mast cells in response to foreign agents and consequently plays a large role in allergic reactions and inflammation. Compounds **1**, **2**, **3**, **4**, and **6** have been illustrated to show histamine release inhibition in rat peritoneal mast cells (210).

Again, many tests have been conducted *in vitro* on ashitaba constituents and their antioxidant and anti-inflammatory effects, but the translatability of these tests to *in vivo* and clinical tests has yet to be determined. Additionally, it should be noted that most substances exhibit some antioxidant effects, especially at high enough concentrations, and calorimetric tests such as those used to evaluate DPPH scavenging activity do not provide strong enough data to confirm antioxidant activity. More robust analyses utilizing cell lines are less likely to yield false positive results and are thus provide more valuable indications of antioxidant capacity.

Antithrombotic activity

Compounds isolated from ashitaba stem tissue show promise as antithrombotic agents due to their antiplatelet activity. Increased levels of plasminogen activator

inhibitor-1 (PAI-1), can result in persistent blood clots leading to thrombotic complications including heart attacks and strokes. TNF-α, a player in inflammation responses, is also involved in the induction of PAI-1 expression. Again, chalcones in ashitaba, namely compounds **2**, **3**, **5**, and **6**, were found to suppress activities induced by TNF-α, resulting in a reversal of PAI-1 increase in human umbilical vein endothelial cells (201).

In another study, **1** and **2** illustrated dose-dependent anti-platelet activity against a number of platelet aggregation inducers, including collagen-, phorbol 12-myristate 13-acetate (PMA), and platelet-activating factor (PAF) in washed rabbit platelets (206). The authors found that **1** and **2** have antiplatelet activity equivalent to aspirin. Because **1** and **2** did not show strong inhibition against thrombin-induced clotting, which is induced through the phospholipase C-β (PLC-β) pathway, the authors concluded that the activity results through the intracellular mobilization of Ca^{2+} by the phospholipase-γ (PLC- γ) pathway, which is also stimulated by collagen and PAF (206).

Blood pressure reducing activity

Although little research has been completed regarding the antihypertensive properties of ashitaba, preliminary research has shown promise for its use in reducing blood pressure. The renin-angiotensin (R-A) system involves the angiotensin I-converting enzyme (ACE), which produces angiotensin II, a vasoconstrictor (202). ACE is a major player in essential hypertension, which is the most prominent type of hypertension diagnosed in the medical field. A compound isolated from ashitaba leaf tissue was found to inhibit ACE from rabbit lung acetone powder. It showed no effect on body weight or

serum lipid levels in spontaneously hypertensive rats (202). Mass spectral data and inhibitory activity data suggested that this compound may be structurally related to nicotianamine. More data is required, both *in vitro* and *in vivo*, to determine the efficacy of ashitaba in treating hypertension.

Antimicrobial activity

Ashitaba chalcones have also shown promise as antimicrobial agents. For example, compounds **1**, **3**, **5**, **7**, **8**, and **13** were found to have potent influenza virus neuraminidase (NA) inhibition on recombinant NA from the 1918 Spanish flu virus (A/Bervig_Mission/1/18), suggesting that they may be useful as anti-influenza agents (181). The authors noted that the activity against NA was influenced by small changes in molecular structure. Elongation of prenyl chains from dimethylallyl groups to geranyl groups caused a two-fold loss of activity. When 2-hydroxy-3-methyl-3-butenyl alkyl (HMB) groups were also prenylated, 2-fold loss of activity was also observed. Conversion of dimethylallyl and geranyl groups to their HMB counterparts, on the other hand, resulted in a gain of activity (181). Compound **5** was found to be the most potent inhibitory agent, and the authors suggested that the location of the HMB group may be responsible for its potency (181).

Compounds **1** and **2** have also been identified as potent antibacterial agents, particularly against Gram-positive bacteria. Using an agar dilution test, these chalcones were shown to have MIC values below 7 μg mL^{-1} for *Staphylococcus aureus* 209-P, and below 2 μg mL^{-1} against *Bacillus subtilis* PCI-219, *B. subtilis* ATCC_6633, *B. cereus* FDA-5, *S. aureus* IFO-3060, *S. epidermidis* IFO-3762, and *Micrococcus luteus* IFO-

12708 (204). These compounds were also shown to have potent antibacterial activity (MIC $\leq$ 1.00 μg mL^{-1}) against plant-pathogenic bacteria, including *Agrobacterium tumefaciens* IFO-3058, *Pseudomonas syringae pv. phaseolicola* IFO-12656, *P. syringae pv. tabaci* IFO-3508, *P. stutzeri* IFO-12510 (204).

Bioavailability

Ashitaba chalcones possess a number of purported health effects, but no reports about the bioavailability of its prenylated chalcones in human tissue currently exist. However, several studies have examined the pharmacokinetic properties of xanthohumol, a prenylated chalcone found in hops, in both humans and rats. Rats and humans given oral administrations of hops typically had nanomolar concentrations of xanthohumol and related prenylflavonoids in their plasma (212-214). In a study conducted on human microbiota-associated rats, the overall excretion of xanthohumol and its related metabolites after two days was only 4.2% of the ingested amount, indicating that this compound is likely hydrolysed by human intestinal microorganisms (213). Additionally, interindividual variability in gut microbiota was found to play a large role in the availability of xanthohumol, and some species of bacteria rapidly hydrolyze this chalcone into 8-prenylnaringenin, a potent phytoestrogen that can affect estrogen signaling pathways (212-214). The associated health effects of the consumption of xanthohumol depends largely on the amount ingested, as well as on the phenotype of the individual ingesting this compound. Whether or not these trends will translate to other prenylated chalcones such as those contained in ashitaba tissue is uncertain, and future studies should aim to determine bioavailability of these compounds. Additionally, studies

determined to identify the *in vivo* differences in metabolism in individuals with variable gut microbiota should be conducted.

Toxicology

The safety of ashitaba was assessed using multiple good laboratory practice (GLP) tests, including a bacterial reverse mutation test, chromosome aberration test, *in vivo* mouse micronucleus test, acute oral toxicity tests, and a 13-week oral toxicity test (215). Additionally, the safety of using ashitaba for cosmetic purposes was assessed using the eye irritancy test (216).

Ashitaba yellow sap chalcone powder was found to be non-mutagenic based on results from the bacterial reverse mutation assay, chromosome aberration assay, and *in vivo* micronucleus assay. Decreased platelet counts were noted in male and female Sprague Dawley rats, which is an expected effect based on known antithrombotic properties of several bioactive chalcones. It was noted that the magnitude of the platelet count reduction is marginal, and not of toxicological significance without other clinical signs (215). Statistically significant levels of serum alkaline phosphatase, total cholesterol, and serum phospholipid and triglycerides were noted in rats fed the highest amount of ashitaba chalcone powder (1000 mg kg^{-1} body weight). This is also an unsurprising discovery based on the known effects of ashitaba on cholesterol transport and lipid metabolism.

Interestingly, male and female rats fed the highest dose showed dilated intestinal lacteals involved in the absorption of dietary fats in the small intestine. Such dilation is indicative of lymphangiectasia, a rare disorder that can lead to edema and its related

complications, including fatigue, abdominal pain, diarrhea, vitamin deficiencies, and weight loss (217). The observation of jejunal lacteal dilation is extremely rare in rodent toxicity studies, so the no observed adverse effect level (NOAEL) of ashitaba powder was concluded to be 300 mg kg^{-1} body weight (215).

To determine the safety of ashitaba as a topical agent, 100 mg of aqueous or ethanol ashitaba leaf extracts were dropped into the eyes of New Zealand White rabbits and the reactions were assessed each day for 7 days. No damages were reported in terms of corneal lesions, turbidity, or eyelid swelling (216). As such, aqueous and ethanol extracts of ashitaba are candidates for use as cosmetic agents.

Although the issue of furanocoumarin toxicity has not been specifically addressed with ashitaba, it should be noted that a number of furanocoumarins have been shown to be phototoxic and photogenotoxic, in addition to interfering with drug metabolism by cytochrome P450 enzymes (218). Ashitaba, as is typical with members of the Apiaceae family, contains bioactive furanocoumarins (**25**) and furanocoumarin analogs containing a tetrahydrofuran, rather than a furan, ring (**23**, **27**). In fact, compound **25** has illustrated phototoxic and photogenotoxic effects in a number of studies (218, 219). A recent report assessed by the Senate Commission on Food Safety determined that compound **25** and its isomer 8-methoxypsoralen are only weakly mutagenic in the absence of UV light, but in the presence of UV radiation, these compounds bind covalently to DNA in bacteria and yeasts, leading to genotoxic and mutagenic effects (219). Because numerous furanocoumarin derivatives are present within ashitaba plant tissue, it is necessary to test individual compounds for phototoxic and photogenotoxic effects. Additionally,

bioavailability is affected both by extract composition as well as the route of administration, and studies are required to determine if phototoxic compounds, such as compound **25**, are at high enough concentrations to be of toxicological concern.

The toxicological data on ashitaba extracts has been addressed to some extent, but more robust toxicological examinations, such as teratogenicity tests, are needed. Additionally, toxicological analyses on isolated compounds should be conducted. In particular, the toxicological profiles of prenylated chalcones (**1**-**22**), the representative structural class of ashitaba, as well as those of furanocoumarins (**23**, **25**, **27**), must be thoroughly characterized to determine ranges of toxicity.

Conclusions

This review summarizes the known phytochemistry and bioactivities of ashitaba. Although there is some inconsistency in the literature, most notably on the effect of ashitaba on adipocyte differentiation, *in vivo* evidence supports the use of ashitaba as a medicinal plant with anti-obesity properties. Although thorough *in vitro* testing has been completed for many of ashitaba's other purported bioactivities, more robust *in vivo* and clinical experiments are needed to confirm the medicinal applications from a clinical standpoint. Clinical testing is warranted to assess ashitaba's anti-diabetic and anti-obesity efficacy, whereas *in vivo* data is needed before pursuing clinical testing for other biological activities. In addition *in vivo* and clinical testing, future studies should focus not only on chalcones **1** and **2**, but also on the bioactivities of other related compounds.

Acknowledgments

Dr. Nicholas Oberlies is acknowledged for his guidance with this project.

CHAPTER III

INTEGRATION OF BIOCHEMOMETRICS AND MOLECULAR NETWORKING TO IDENTIFY BIOACTIVE CONSTITUENTS OF ASHITABA

This chapter has been published in the journal Planta Medica and is presented in that style. Caesar, L.K., Kellogg, J.J., Kvalheim, O.M., Cech, R.A., Cech, N.B. Planta Medica. 2018, 84(9-10), 721-728.

Caesar, L.K. performed and interpreted all antimicrobial assays, completed all chromatographic separations, conducted mass spectral analysis, produced molecular networks, created all the figures, and wrote the manuscript. Kellogg, J.J. and Kvalheim, O.M. worked with Caesar, L.K. to produce selectivity ratio plots and complete statistical analysis. Cech, R.A. grew the plants from which compounds were isolated, and confirmed the identity of the *Angelica keiskei* botanical specimen. Cech, N.B. provided suggestions and edits throughout manuscript preparation.

Introduction

The complexity of botanicals makes them a rich source for medicinally useful compounds, but leads to many analytical challenges. The traditional workflow for natural product discovery is bioassay-guided fractionation (220, 221), in which bioactive extracts and subsequent fractions are chromatographically separated and retested for bioactivity until active compounds have been isolated. Because botanical extracts contain thousands of individual constituents, it is often difficult to assign activity to individual components, thus, the most abundant or easily isolatable compounds are often presumed to be responsible for bioactivity (131, 132). New methods are needed that will enable isolation

efforts to be focused on those components most likely to be responsible for the desired biological activity.

Compounds from nature have been utilized to treat microbial infections throughout history (222), and some sources estimate that up to two-thirds of antibacterial agents on today's market are derived from natural products (223). The virtually limitless chemical diversity of natural products, particularly botanicals, results from their complex biosynthetic pathways, and many plant secondary metabolites, including flavonoids, alkaloids, and coumarins, have shown antimicrobial activity (221, 222, 224-227). *Angelica keiskei* Koidzumi (Apiaceae), or ashitaba, is a member of the *Angelica* genus native to the southernmost islands of Japan that is popularly utilized as a food and a medicinal herb, purportedly to extend life expectancy, increase vitality, and to treat a broad range of diseases and infections. Most of these activities result from the action of unique prenylated chalcones, as well as coumarins and flavanones (reviewed in (228)). Two compounds from *A. keiskei,* 4-hydroxyderricin (**1**) and xanthoangelol (**2**) have been shown to possess activity against methicillin-resistant *Staphyloccoccus aureus* (MRSA) (204). Additionally, *A. keiskei* chalcones xanthoangelol F and isobavachalcone are active against other Gram-positive organisms, though they have not been tested against pathogenic bacteria such as MRSA (229). With this study, we sought to employ antimicrobial extracts of *A. keiskei* as a test case for the development of new methods to prioritize bioactive compounds early in the isolation process for a complex botanical.

In combination with chromatographic techniques, mass spectrometry can be utilized to analyze hundreds of secondary metabolites simultaneously (132, 230, 231).

Using a process called biochemometrics, quantitative chemical information and biological activity data can be incorporated into a statistical model. With this statistical modeling approach, it is possible to discovery chemical patterns related to bioactivity (132). Partial least squares (PLS) analysis can be used in combination with chromatographic and mass spectrometric data to correlate metabolite profiles with biological data (232). A recent study from our laboratory showed that selectivity ratio analysis was useful for the identification of trace bioactive constituents in fungal extracts without being confounded by highly abundant compounds (132). The selectivity ratio compares the correlation and covariance to the residual variance, and provides a quantitative measurement of the ability of a given variable to differentiate between active and inactive groups (233).

Biochemometric analysis is helpful for distinguishing between active and inactive chemical constituents, but it is also useful to obtain structural information for the purpose of prioritizing new compounds for isolation. To address this, we have utilized the Global Natural Product Social Molecular Networking (GNPS) database (143)to build molecular networks from mass spectral fragmentation data. These fragmentation data provide useful chemical information, and structurally similar molecules should possess similar mass spectral fragmentation patterns. By comparing cosine similarity scores of individual compounds' fragmentation patterns, GNPS can produce visual networks comprised of chemically related compounds and enables the identification of known compounds, molecular families, and structural analogs. By combining GNPS networking with biochemometric analysis, we propose that it would be possible to identify the structural

classes of putative active molecules. The goal of this project is to utilize this integrated approach to prioritize isolation efforts on biologically relevant compounds from *A. keiskei* and to gain a more comprehensive understanding of which constituents contribute to the antimicrobial activity of this botanical against MRSA.

Results and Discussion

The first goal of this study was to utilize biochemometric analysis to identify putative bioactive constituents contributing to the antimicrobial activity of *A. keiskei*. Bioactivity screening demonstrated complete inhibition of methicillin-resistant *S. aureus* (MRSA, strain USA300 LAC strain AH1263) (234) by the *A. keiskei* extract at 10 µg/mL. This extract was then fractionated in several stages (see fractionation schemes, Appendix C, Figures S1 and S2), with the fractions displaying the most pronounced antimicrobial activity against MRSA prioritized for further isolation (Table 5).

Bioactivity and mass spectral data from the second stage of fractionation (AK-3-1 through AK-3-8 and AK-4-1 through AK-4-4) (Table 5) were utilized to produce a biochemometric model predicting which constituents were responsible for antimicrobial activity. The internally cross-validated model generated five components that accounted for 83.61% of the independent (mass spectral), and 99.93% of the dependent (growth inhibition) variation (component 1: 32.58% independent, 53.16% dependent; component 2: 24.85% independent, 30.29% dependent; component 3: 11.54%, 13.98%; component 4: 7.86%, 1.93%; component 5: 6.79%, 0.57%).

Table 5. Antimicrobial Activity of *Angelica keiskei* Koidzumi (AK) Crude Extract (CR) and Second-Stage Fractions AK-3-1 through AK-4-4[a]

Sample	Methicillin-resistant *S. aureus* growth inhibition (%)	
	50 µg/mL	**5 µg/mL**
Chloramphenicol[b]	100 ± 0	46.7 ± 1.8
AK-CR	99.22 ± 0.39	6.4 ± 6.0
AK-3-1	0 ± 0[b]	21 ± 16
AK-3-2	99.35 ± 0.65	26.0 ± 1.3
AK-3-3	99.09 ± 0.91	11.14 ± 0.79
AK-3-4	100 ± 0	0 ± 0
AK-3-5	90.7 ± 3.3	99.61 ± 0.23
AK-3-6	0 ± 0[b]	26 ± 15
AK-3-7	0 ± 0	0 ± 0
AK-3-8	0 ± 0	0 ± 0
AK-4-1	97.4 ± 2.4	19.76 ± 0.26
AK-4-2	98.8 ± 1.2	98.95 ± 0.47
AK-4-3	99.74 ± 0.26	3.2 ± 1.2
AK-4-4	0 ± 0	0.66 ± 0.66

[a] Growth inhibition of methicillin-resistant *S. aureus* strain (MRSA USA300 LAC strain AH1263) (234) relative to vehicle control measured turbidimetrically by OD600. Data presented are the result of triplicate analyses ± SEM. [b] Chloramphenicol (Sigma-Aldrich, 98% purity) served as the positive control.
[b] Higher concentration samples of AK-3-1 and AK-3-6 show lower activity than their low-concentration counterparts, likely due to low solubility in aqueous media at high concentrations.

To interpret the model and tentatively identify the chemical entities responsible for the MRSA growth inhibition, a selectivity ratio plot was generated (Figure 15A). This plot revealed several marker ions that were strongly correlated with bioactivity, but could not provide structural information about these components. To generate such structural information, molecular networks were generated using MS/MS data from second-stage and third-stage chromatographic fractions (fractions resulting from two or three rounds of chromatographic separation, Appendix C, Figure Fig. S1). The resulting molecular networks were filtered using the biochemometric selectivity scores to identify molecular families of putative active compounds and assign tentative structures to candidate molecules (Supplementary Fig. S3). Interestingly, one second-stage molecular network

and one third-stage molecular network identified the chalcones 4-hydroxyderricin (**1**) and xanthoangelol (**2**), the only known anti-MRSA compounds from *A. keiskei* (204).

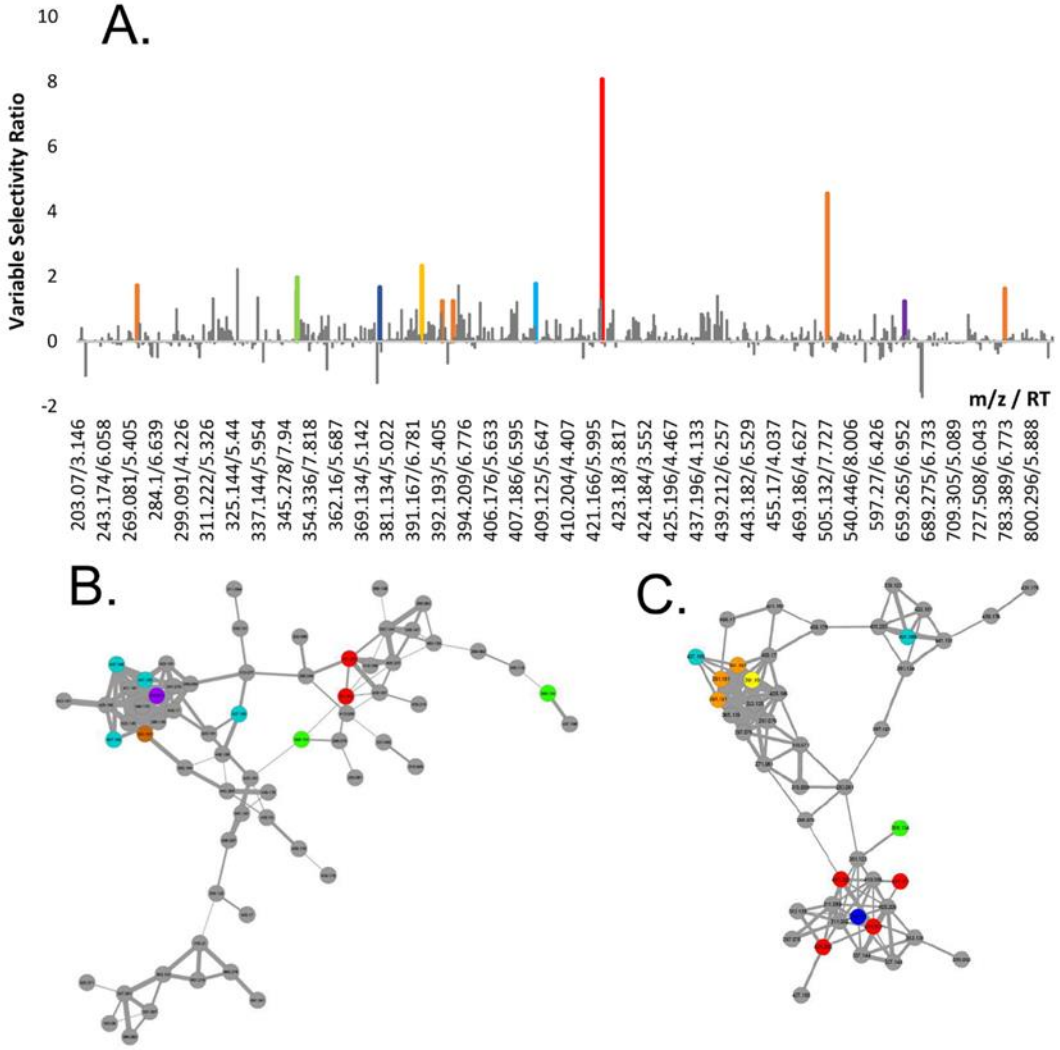


Figure 15. Selectivity Plot (A) and Selected Molecular Networks of Second-Stage (B) and Third-Stage (C) Fractions of *A. keiskei* Root Extract. Bars have been color coded in A and points have been color coded in B and C only if they were both correlated with bioactivity and appeared in molecular networks of interest. Predicted active compounds in A appeared almost exclusively in these networks, indicating that a particular class of compounds is responsible for *A. keiskei*'s antimicrobial activity.

Other known *A. keiskei* chalcones were also identified (Figure 16). The same networks also contained masses of seven of the top ten contributors to bioactivity (marker ions A-G, Table 6) based on the biochemometric model (Figures 15B-15C), suggesting that chalcones are responsible for *A. keiskei's* antimicrobial efficacy against MRSA. The combination of biochemometrics and molecular networking enabled identification of a subset of these chalcones for prioritization and subsequent analysis, making it possible to predict the identity of *biologically active* extract components prior to isolating them.

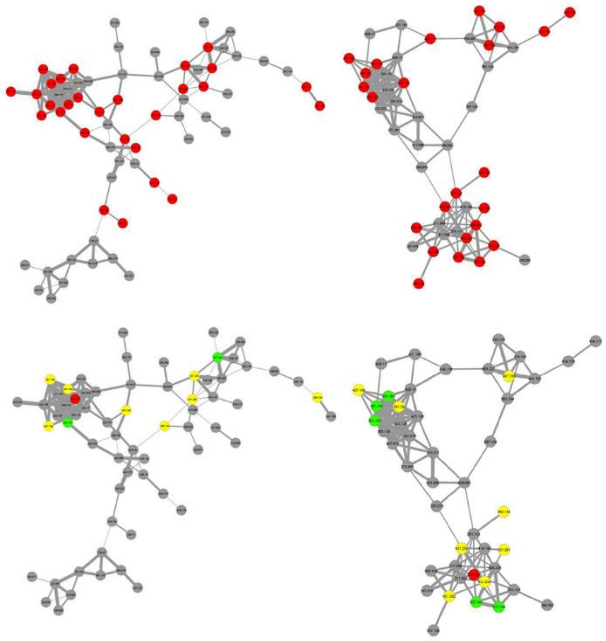


Figure 16. Molecular Networks Comprised of Compounds Detected in *A. keiskei* Built from Fractions Following One (Left) and Two (Right) Stages of Fractionation. In top networks, compounds marked in red match accurate masses of known *A. keiskei* chalcones. In bottom networks, green compounds match accurate masses of known antimicrobials **1** and **2**, yellow compounds match known chalcones that have not been shown to possess anti-MRSA activity, and red compounds were correlated with bioactivity based on biochemometric selectivity ratio analysis but do not match known masses from the literature.

Table 6. Tentative Identification of Putative Bioactive Chalcones from *A. keiskei*.

Marker ion	Ion/retention time (molecular formula, δ (ppm))	Adducts and fragments (molecular formula, δ (ppm))	Tentative identification(s)
A	421.202 $[M-H]^-$ / 6.23 ($C_{26}H_{29}O_5^-$, 1.189)		4,2',4'-trihydroxy-3'-[(2*E*, 5*E*)-7-methoxy-3,7-dimethyl-2,5-octadienyl]chalcone [a] Xanthoangelol G [a]
B	391.191 $[M-H]^-$ / 6.77 ($C_{25}H_{27}O_4^-$, 0.168)	505.184 $[M-H + TFA]^-$ ($C_{25}H_{27}O_4 + C_2HF_3O_2$, 0.399) 271.134 $[M-H - C_8H_8O]^-$ ($C_{17}H_{19}O_3^-$, 2.141) 783.389 $[2M-H]^-$ ($2C_{25}H_{28}O_4 - H$, 0.886)	Xanthoangelol [b]
C	391.191 $[M-H]^-$ / 5.59 ($C_{25}H_{27}O_4^-$, 0.168)		Xanthoangelol I [a]
D	351.123 $[M-H]^-$ / 5.52 ($C_{21}H_{19}O_5^-$, 0.708)		Xanthoangelol K [b]
E	407.186 $[M-H]^-$ / 6.58 ($C_{25}H_{27}O_5^-$,0.371)		Xanthoangelol B [a] (2E)-1-[3,5-dihydroxy-2-methyl-2-(4-methyl-3-[penten-1-yl)-3,4-dihydroxy-2H-chromen-8-yl]-3-(4-hydroxyphenyl-2-propen-1-one) [a] (2E)-1-[4-hydroxy-2-(2-hydroxy-6-methyl-5-hypten-2-yl)-2,3-dihydro-1-benzofuran-5-yl]-3-(4-hydroxyphenyl)-2-propen-1-one [a]
F	379.155 $[M-H]^-$ / 5.97 ($C_{23}H_{23}O_5^-$, 1.19)		Potentially new chalcone derivative [c]
G	439.211 $[M-H]^-$ / 5.17 ($C_{26}H_{31}O_6^-$, 2.422)		Potentially new chalcone derivative [c]

[a] previously reported from *Angelica keiskei* Koidzumi, identified using accurate mass data (228).
[b] isolated and confirmed by NMR
[c] accurate masses do not match accurate masses of known *A. keiskei* chalcones, yet these masses appeared in chalcone molecular networks, indicating that they may be new chalcone derivatives.

Fifteen of the features in networks of interest matched the reported accurate masses of known chalcones (228) that have not yet been associated with antimicrobial

activity (Figure 16). Of these, five were predicted as potentially contributing to bioactivity by the biochemometric model, including the top contributor at *m/z* 421.202. Two additional compounds in these networks were identified among the top ten contributors by the biochemometric model that did not match accurate masses of bioactive chalcones from *A. keiskei* (Figure 16). Because these compounds clustered with known chalcones based on similarities in mass spectral fragmentation patterns (Figure 16), it was predicted that other chalcone antimicrobials might be present.

Biochemometric and molecular networking analysis identified marker ions associated with activity (Table 6). Purification of active *A. keiskei* fractions was conducted to assess the predictive accuracy of this approach, and four compounds were isolated (Figure 17). The two known anti-MRSA compounds from *A. keiskei*, **1** and **2**, were isolated using a combination of normal- and reversed-phase chromatography. Compound **1** was isolated at 98% purity following two stages of normal-phase flash chromatography and one stage of reversed-phase flash chromatography. Compound **2** was obtained at 95% purity following three stages of fractionation using both normal-phase flash chromatography and reversed-phase preparative-scale HPLC. The structures of compounds **1** and **2** were confirmed with ^{1}H and ^{13}C NMR by comparing to literature data (235) (Appendix C, Figures S4-S7).

Two additional chalcones, **3** and **4**, were isolated following a scale-up extraction and isolation. Compound **3** was isolated with 96% purity following two rounds of normal-phase flash chromatography, and **4** at 99% purity required an additional round of reversed-phase preparative HPLC. ^{1}H and ^{13}C NMR were utilized to confirm the

identities of these compounds by comparing to published data (178, 180) (Appendix C, Figures S8-S12). For **4**, HMBC data were collected to confirm the presence of a ketone peak that did not appear in the ^{13}C NMR spectra (Appendix C, Figure S12), likely due to keto-enol tautomerization.

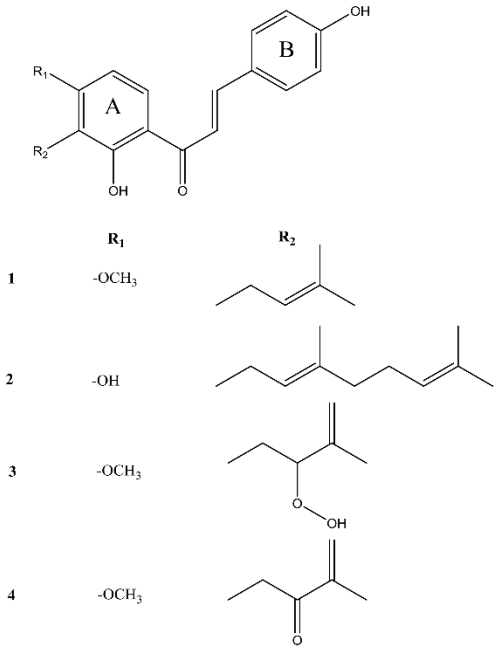


Figure 17. Structures of Compounds 1-4, which were Isolated from Ashitaba (*Angelica keiskei*) and Assessed for Antimicrobial Activity.

By integrating biochemometrics and molecular networking into the traditional bioassay-guided fractionation workflow, it was possible to prioritize minor constituents in *A. keiskei* for isolation (see workflow, Appendix C, Figure S3). Using biochemometrics to filter molecular networks and focus on specific structural classes, a subset of chalcone derivatives were identified that were most likely to possess antimicrobial activity and were prioritized for isolation. With this method, known,

abundant antimicrobial compounds **1** and **2** were isolated, similar to previous bioassay-guided fractionation approaches alone (204). Compounds **1** and **2** demonstrated MICs against MRSA (USA300 LAC strain AH1263) (234) of 4.6 µM and 4.0 µM, respectively (Table 7, Figure S13). The biochemometrics/GNPS approach also enabled isolation of an additional low abundance antimicrobial compound (**4**), marker ion D (Table 6) that has not previously been reported to possess antimicrobial activity. In selectivity ratio plots, **4** was listed as the fourth top contributor to the observed biological activity of *A. keiskei* despite its low relative abundance (Figure 15A, Table 6). In the base-peak chromatogram of the *A. keiskei* root extract, the peak area associated with **4** only accounted for 0.8% of the total fraction (Figure 18). Compound **4** did inhibit growth of MRSA (IC_{50} at 168 µM, Table 7, Appendix C, Figure S13) although did not reach MIC at the highest concentration tested (284 µM). Finally, as additional confirmation, we also isolated **3**, which appeared in the chalcone molecular network (Figure 16) but was not predicted to be antimicrobial. As predicted by biochemometrics, **3** did not possess antimicrobial activity, despite structural similarity to active compounds. Collectively, the agreement between predicted and observed biological activity of **1-4** demonstrates that the biochemometrics process as employed can be effective for identifying a subset of molecules for isolation based on their likely biological activity.

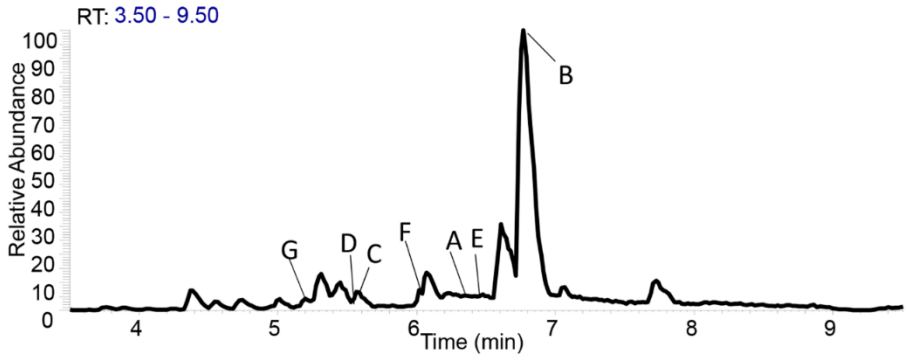


Figure 18. Base-Peak Chromatogram of Ethyl Acetate *A. keiskei* Root Extract with Peaks of Interest Identified by Biochemometric Selectivity Ratio Analysis. This analysis was successful in enabling prioritization of trace peaks of interest for isolation.

Table 7. MIC and IC_{50} Data for Compounds 1-4 against Methicillin-Resistant *S. aureus* (MRSA USA300 LAC Strain AH1263) Relative to Vehicle Control Measured Turbidimetrically by OD_{600}. Presented data were calculated using four-parameter logistic curves of triplicate data.

Compound	MIC[a]	IC_{50}
1	4.6 µM	2.0 µM
2	4.0 µM	2.3 µM
3	--	--
4	--	168 µM

[a] The MIC value expressed is likely higher than the actual MIC value, which lies somewhere between the lowest tested concentration that inhibited bacterial growth and the highest tested concentration that did not completely inhibit bacterial growth (236).

The results described here are consistent with previous studies which suggest that prenyl- and geranyl- moieties on the A-ring of chalcones (present in **1** and **2**) are associated with antimicrobial activity (229). Compound **3** has a markedly different side chain from **1** and **2,** with a flexible peroxide group, which is likely responsible for its decreased antimicrobial activity. Compound **4,** though it does not contain an prenyl side chain, could possess weak activity due to the similarity of its side chain in rigidity and size to the prenyl substituent seen in **1**.

Several additional features identified as possibly contributing to biological activity were identified in GNPS as chemically related to isolated chalcones **1-4** (Figures 15 and 16). Based on these networks and accurate mass data of these compounds, we tentatively identified these compounds (Table 6). Unfortunately, material was too limited to isolate these compounds or assess biological activity. From a drug discovery standpoint, however, this approach is useful in dereplication, as it allowed us to identify these compounds as chalcones early in the fractionation process. Since chalcones are well documented antimicrobials (237), we did not complete an additional scale up to pursue their isolation.

In this example, marker ion A (Table 6) at *m/z* $[M-H]^-$ 421.202 eluting at 6.2 minutes was identified as the constituent most correlated with bioactivity and accounted for 0.4% of the total extract based on peak area. Unfortunately, even with a scale up extraction and chromatographic efforts tailored to this specific compound, isolation efforts for this compound were unsuccessful. This failure to isolate the active constituent demonstrates one of the inherent limitations of the biochemometric approach for identifying bioactive compounds. While it is possible based on mass spectrometric data to identify minor compounds that may have important biological activity, it may be infeasible (due to limited quantity) to isolate such minor compounds for confirmation of structure and activity.

One limitation of this study is that biochemometric analysis did not predict biological activity for the most abundant isotopes of **1** and **2**, despite the confirmed antimicrobial activity of these compounds. Based on relative peak area, **2** accounted for

37.8% of the relative abundance in the EtOAc extract, and **1** accounted for 12.5%. The high abundance and antimicrobial potency of these compounds likely led to a mismatch in biological and chemical data. While the relative peak area of these compounds varied in every fraction under study, the biological activity was saturated at 100% in multiple fractions. Consequently, the linearity between the relative abundance of these compounds and their corresponding bioactivity was likely skewed, leading to false negative results. Although the $[M-H]^-$ peak for the most abundant (^{12}C) isotope of **2** was not identified as active (Figure 15), several of the ^{13}C isotopes as well as the TFA adduct, and an in-source fragment of this compound were predicted to be active (marker ion B, Table 6). The adducts and isotopes of **2** were only evident in fractions where **2** was extremely abundant, and consequently, they were identified by the selectivity score as marker ions related to bioactivity. The identification of an active isotope of a compound that is not itself predicted to be active is clearly an artifact of an error in the data analysis process, given that all isotopes co-occur in the sample, and the adducts are formed in the ionization process and likely not present in the sample at all.

An important goal for the comprehensive characterization of a botanical medicine should be to isolate minor constituents within the extract. However, it is not feasible to isolate all minor constituents in a complex mixture, so putative bioactive constituents must be prioritized. A major strength of the biochemometric selectivity ratio analysis is its ability to identify low-abundance constituents contributing to activity without being confounded by compounds of high abundance. However, this strength comes with an important weakness in that bioactive compounds of high abundance may be overlooked.

This weakness can easily be overcome, however, if this statistical analysis is incorporated into the traditional bioassay-guided fractionation workflow, which favors the isolation of abundant active compounds. It is also possible that this limitation could be addressed by diluting samples to reduce the level of high abundance compounds, although this approach would come at the expense of sacrificing response of those present at low abundance.

In combination with bioassay-guided fractionation, biochemometrics and molecular networking can be utilized to identify structural families of putative active constituents present at very low levels, allowing for the prioritization of isolation of both high and low abundance components that contribute to activity, or alternately, enabling the dereplication of known bioactive compounds and their structural analogs. The latter application is important because it prevents time being wasted on reisolating known active compounds. Had we been searching for bioactive compounds with novel structures only, we may have chosen not to pursue further isolation with *A. keiski* once we identified chalcones as the major class of active constituents within this plant. However, for the purpose of this study, a botanical containing known antimicrobial constituents served as a useful test case. The approach employed here not only facilitated the identification of a trace antimicrobial constituent from *A. keiskei*, but also yielded new and more complete information about which constituents are responsible for the antimicrobial activity of this botanical. Additionally, it provided insight into which structural characteristics of chalcones are associated with their antimicrobial effects.

Materials and Methods

General experimental procedures

NMR spectra were obtained using a JEOL ECA-500 MHz spectrometer. UPLC-MS analysis was completed in both negative and positive modes using an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific) connected to an Acquity UPLC system (Waters Corporation). When collecting UPLC-MS data, 3 μL of 1 mg/mL samples suspended in MeOH were injected into the column. Using a flow rate of 0.3 mL/min, samples eluted from the column (BEH C18 1.7 μm, 2.1 x 50 mm, Waters Corporation) using the following gradient with solvent A consisting of water with 0.1% formic acid and solvent B consisting of acetonitrile with 0.1% formic acid: 90:10 (A:B) from 0-0.5 min, increasing to 0:100 (A:B) from 0.5-8.0 min. The gradient was held at 100% B for 0.5 min, before returning to starting conditions over 0.5 min and held from 9.0-10.0 min. Mass analysis was completed in both positive and negative ionization modes over a scan range of 150-2000 with the following settings: capillary voltage at -21.00 V, capillary temperature at 275.00 °C, tube lens offset at -95.00 V, spray voltage at 3.50 kV, sheath gas flow at 30.00, and auxiliary gas flow at 15.00. The top 4 most intense ions were fragmented with CID set to 35.0.

Flash chromatographic separations were completed using a CombiFlash RF system (Teledyne-Isco) and examined using a PDA detector and an evaporative light scattering detector (ELSD). Preparative and analytical HPLC separations were conducted with a Varian HPLC system (Agilent Technologies) using Galaxie Chromatography

Workstation software (version 1.9.3.2, Agilent Technologies). All chemicals were acquired through Sigma-Aldrich and were spectroscopic or microbiological grade.

Plant material

Fresh roots of *Angelica keiskei* Koidzumi were collected on November 14, 2015 from Strictly Medicinal Seeds in Williams, Oregon (Sample # 12421, N 42°12'17.211", W 123°19'34.60). Scale up material was completed using plant material from the same source collected on December 29, 2016 (Sample #12444, N 42°12'17.211", W 123°19'34.60). The identity of this plant material was confirmed by Richard A. Cech at Strictly Medicinal Seeds, and a voucher specimen was deposited at the herbarium of the University of North Carolina at Chapel Hill (NCU627665).

Extraction

Fresh *A. keiskei* roots were dried in a single-wall transite oven (Blue M Electric Company) at 40 °C for 24 hours. The resulting dry mass (138.90 g) was ground using a Wiley Mill Standard Model No. 3 (Arthur Thomas Company) and submerged in MeOH at 160 g/L for 24 hours three times. The resulting MeOH extract was concentrated *in vacuo* and then subjected to liquid-liquid extraction. First, the extract was defatted by partitioning between 10% aqueous MeOH and hexane (1:1). The dried aqueous MeOH layer was partitioned further between 4:5:1 EtOAc/MeOH/H_2O. To remove hydrosoluble tannins, the EtOAc layer was washed with a 1% NaCl solution. The resulting EtOAc extract was dried under nitrogen, yielding 3,650.32 mg dried extract, before further experimentation. Scale up material (964 g) was dried, extracted, and partitioned using the

same methods listed above, ultimately yielding 18.10 g of dried EtOAc extract for subsequent chromatographic separation.

Chromatographic separation and isolation

The isolation scheme is provided in Appendix C (Figures S1 and S2). The first-stage separations of the EtOAc extract (3,100 mg) were conducted using normal-stage flash chromatography (40 g silica gel column) at a 40 mL/min flow rate with a 35 min hexane/$CHCl_3$/MeOH gradient. The last two fractions (AK-3 and AK-4) were subjected to a second stage of normal-phase flash chromatography. Fraction 3 (AK-3, 1355 mg) was separated again with a 40 g silica gel column at a flow rate of 40 mL/min and fraction 4 (AK-4, 536 mg) was separated on a 12 g silica column with a flow rate of 30 mL/min. Each run lasted 45 minutes, and was completed using a hexane/EtOAc/MeOH gradient. The most active fraction from the separation of AK-3 (fraction 2, AK-3-2, 1000 mg) was subjected to a final round of reversed-phase flash chromatography using a 130g C18 reversed phase RediSep Rf column with an 85 mL/min flow rate. A 25-minute gradient of CH_3CN/H_2O was used, starting at 40:60 and increasing to 85:15. It was increased to 100:0 for 5 minutes, upon which starting conditions were re-established. Compound **1** eluted at 18 min (234.45 mg, 98% purity, 7.6% yield). Fraction AK-4-2 (364 mg) was also subjected to a final round of reversed-phase preparative HPLC injected onto a Luna preparatory column (5 µm PFP, 250 × 21.20mm; Phenomenex). The 35 minute run began at 40:60 $CH_3CN:H_2O$ and was increased to 100:0 over thirty minutes. Compound **2** was collected from 28-35 minutes (284.59 mg, 95% purity, 9.1% yield).

Compounds **3** and **4** were isolated following scale up extraction. First, 17.5 g of EtOAc extract were separated on a 120g silica column with an 85mL/min flow rate using the same hexane/$CHCl_3$/MeOH gradient as used for the first fractionation of original extract. The second fraction (S-AK-2, 5.3 g) was separated again using normal-phase flash chromatography on a 120 g silica column at 85 mL/min flow rate with a 55-minute gradient of hexane/EtOAc/MeOH. Compound **3** eluted at 31 minutes (150 mg, 96% purity, 0.85% yield). Fraction 4 (S-AK-2-4, 172 mg) was subjected to a final 45-minute round of reversed-phase preparative HPLC on a Gemini-NX preparatory column (5 µm C18, 250 × 21.20 mm; Phenomenex) at a flow rate of 21.4 mL/min with a gradient of 55:45 CH_3CN:H_2O. Compound **4** (1.5 mg, 99% purity, 0.0086% yield) eluted at 19 minutes.

*4-hydroxyderricin (**1**):* yellow crystalline solid; HRESIMS *m/z* 337.1438 $[M-H]^-$ (calculated for $C_{21}H_{21}O_4^-$, 337.1440); ^1H NMR (500 MHz, $CDCl_3$) and ^{13}C NMR (125 MHz, $CDCl_3$) chemical shifts matched literature values (205) and are provided in Appendix C (Figures S4 and S5).

*Xanthoangelol (**2**):* yellow crystalline solid; HRESIMS *m/z* 391.1907 $[M-H]^-$ (calculated for $C_{25}H_{27}O_4^-$, 391.1909); ^1H NMR (500 MHz, $CDCl_3$) and ^{13}C NMR (125 MHz, $CDCl_3$) chemical shifts matched literature values (205) and are provided in Appendix C (Figures S6 and S7).

*Xanthoangelol E (**3**):* yellow, amorphous powder; HRESIMS *m/z* 369.1340 $[M-H]^-$ (calculated for $C_{21}H_{21}O_6^-$, 369.1338); ^1H NMR (500 MHz, DMSO) and ^{13}C NMR (125

MHz, DMSO) chemical shifts matched literature values (178) and are provided in Appendix C (Figures S8 and S9).

*Xanthoangelol K (**4**):* yellow amorphous powder; HRESIMS *m/z* 351.1231 $[M-H]^-$ (calculated for $C_{21}H_{19}O_5^-$, 351.1232); ^1H NMR (500 MHz, $CDCl_3$), ^{13}C NMR (125 MHz, $CDCl_3$), and HMBC (400 MHz, $CDCl_3$) chemical shifts matched literature values (180) and are provided in Appendix C (Figures S10-S12).

Antimicrobial assay

Antimicrobial activity was monitored by assessing growth inhibition of a laboratory strain of *Staphylococcus aureus* (SA1199) (238) and a clinically relevant strain of methicillin-resistant *S. aureus* (MRSA USA300 LAC strain AH1263) (234), obtained from Dr. Alexander Horswill at the University of Colorado Anschutz Medical Campus. Cultures were grown from a single colony isolate of each strain in Müeller-Hinton broth (MHB) and diluted to 1.0 x 10^5 CFU/mL based on absorbance at 600 nm (OD_{600}).

Samples were screened in triplicate at final concentrations of 10 and 100 μg/mL or 5 and 50 μg/mL. Samples were dissolved in 1:1 EtOH/DMSO (v/v) and diluted with MHB to prepare final concentrations in broth with less than 2% EtOH/DMSO. The known antibiotic chloramphenicol (98% purity, Sigma-Aldrich) was used as a positive control at the same concentrations as tested extracts. The vehicle was 2% EtOH/DMSO in MHB. Each well was inoculated with bacteria and incubated for 24 hours at 37 °C. OD_{600} was evaluated after incubation and used to calculate the percent growth inhibition.

All fractions were subjected to analysis and active fractions were chosen for further fractionation.

Minimal inhibitory concentrations (MICs) were calculated for pure compounds based on the Clinical Laboratory Standards Institute (CLSI) standard protocols (236). Isolated compounds or chloramphenicol (positive control, 98% purity, Sigma-Aldrich) were added to 96-well plates in triplicate at concentrations ranging from 0-100 µg/mL in MHB. Broth containing 2% 1:1 EtOH/DMSO was used as the vehicle control. The concentration of EtOH/DMSO was set at a fixed value of 2% for all wells. After a 24-hour incubation at 37 °C, OD_{600} values were measured using a Synergy H1 microplate reader (Biotek). The MIC was defined as the concentration at which no statistically significant difference between the blank wells (containing sample and broth but no bacteria) and the treated sample was observed.

Biochemometric analysis

LC-MS data were collected in both negative mode and positive mode and individually analyzed, aligned, and filtered utilizing MZmine 2.21.2 (http://mzmine.sourceforge.net/) (239). Raw mass spectral data files from second-stage fractions were uploaded for peak picking into MZmine based upon *m/z* values within each spectrum above a set baseline for all batch samples. Chromatograms were constructed for all *m/z* values lasting longer than 0.1 minute, following which they were deconvoluted using algorithms that were applied to chromatograms to recognize individual peaks. The peak detection parameters were set as follows: noise level (absolute value) at 1.25×10^6 (positive mode) and at 2×10^6 (negative mode), minimum peak

duration at 0.5 seconds, *m/z* variation tolerance at 0.05, and *m/z* intensity variation tolerance at 20%. Peaks were aligned if their masses were within 5 ppm and their retention times were a maximum of 0.15 minutes from one another. Peak list filtering and retention time alignment were completed to produce an aligned peak list. The resulting data matrix, consisting of *m/z*, retention time, and peak area, was imported into Excel (Microsoft) and merged with bioactivity data from samples at tested at 5 μg/mL to form the final data set for biochemometric analysis.

Biochemometric analysis was completed using Sirius version 10.0 statistical software (Pattern Recognition Systems) (240). Before analysis, data were adjusted using a fourth root transformation to normalize noise across treatments (241). An internally cross-validated PLS model was then produced using 100 iterations and a significance level of 0.05. Statistical algorithms internal to the Sirius software utilized model predictions to produce selectivity ratios identifying putative antimicrobial constituents.

Molecular networking analysis

Mass spectral data were converted to mzXML format using FileZilla version 3.14.1, part of the ProteoWizard platform (http://proteowizard.sourceforget.net/#). Following file conversion, mass spectral and fragmentation data were uploaded to the GNPS data analysis portal in 3 groups, where fractions active at 5 μg/mL were included in Group 1, fractions active at 50 μg/mL were included in Group 2, and inactive fractions were included in Group 3. These data were then combined into consensus spectra using the MS-clustering algorithm (242) within the Global Natural Product Social Molecular Networking (GNPS) database (143).

Molecular networks were produced using the online GNPS workflow. First, MS/MS peaks within 17 Da of the precursor *m/z* were removed, and only the top six fragment peaks were compared for analysis. Using MS-Cluster, consensus spectra were produced with a parent mass tolerance of 0.5 Da and an MS/MS fragment ion tolerance of 0.3 Da. Consensus spectra containing fewer than 10 spectra were discarded. Molecular networks were subsequently produced, and compounds were connected if they had a cosine score (similarity score) above 0.65 and more than 6 matched fragment peaks. If more than 10 compounds shared a cosine score above this threshold with a given compound, only the top 10 most similar compounds were connected. Parameters for third-stage fractions were the same, except that the minimum cluster size was adjusted to 100. Fragmentation patterns were compared to databases within GNPS, including the GNPS Library, the GNPS-NIH-Natural Products Library, GNPS Prestwick Phytochemical Library, and the RESPECT Library to tentatively ID components matching MS/MS patterns already contained within the system. Networks were viewed in GNPS using the network visualizer in addition to being imported to Cytoscape (243) for visualization. To simplify investigation of networks, nodes containing accurate masses identified by biochemometric analysis as putative active compounds were prioritized for structural characterization.

Acknowledgements

This research was supported by the National Center for Complementary and Integrative Health of the National Institutes of Health under award numbers 5 T32 AT008938, and 1R01 AT006860. The authors would also like to thank Dr. Laura

Sanchez for her help with molecular networking, and Dr. Alexander Horswill for his provision of microbial strains utilized for this project.

CHAPTER IV

HIERARCHICAL CLUSTER ANALYSIS OF TECHNICAL REPLICATES TO IDENTIFY INTERFERENTS IN UNTARGETED MASS SPECTROMETRY METABOLOMICS

This chapter has been published in the journal Analytica Chimica Acta and is presented in that style. Caesar, L.K., Kvalheim, O.M., Cech, N.B. Anal Chim Acta. 2018, 1021, 69-77.

Caesar, L.K. completed all chromatographic separations and collected all mass spectrometric data, conducted hierarchical cluster analysis, developed the relative variance cutoff described herein, and wrote the manuscript and prepared all figures. Kvalheim, O.M. assisted with hierarchical clustering analysis and provided guidance for statistical analysis sections. Cech, N.B. provided edits and suggestions throughout manuscript preparation.

Introduction

Metabolomics is a growing field in which analysts seek to comprehensively analyze and compare quantities of metabolites (small molecules) in biological samples (186, 244-249). Creative applications of metabolomics span over a wide range of subjects, and this tool has been applied to facilitate the understanding of disease pathogenesis (247), to study the effects of diet and drug interactions (250), for biomarker identification (251-253), and for natural products drug discovery (132, 248, 254). Mass spectrometry is often the analytical technology of choice for metabolomics research, due to its unparalleled ability to detect metabolites present at low levels (244). The large data

sets generated using untargeted mass spectrometry metabolomics can, however, be difficult to deconvolute.

Because metabolites are not directly coded in to an organism's DNA and are often influenced by stage of life, source of material, and environmental conditions, it is quite difficult to define the number of metabolites in a given biological sample (255). As such, a central challenge in the metabolomics field is data analysis (245, 249, 256, 257). The data sets generated from metabolomics analysis may contain tens of thousands of individually detected compounds, which may include many experimental artefacts (246, 258). The effective handling of such large data sets is a unique challenge (244), and investigation into these data sets requires advanced statistical tools capable of extracting relevant information from the vast quantity of data produced (256). Unfortunately, these multivariate techniques are often used incorrectly and lack proper validation (256). False positives are likely to occur when performing statistical analyses on these types of data sets (244) since the number of samples analyzed is typically much lower than the number of variables analyzed. As a consequence, overfitting the data is a serious concern (246, 256). The problem of false positives grows when peaks not associated with the sample are included in the dataset, and effective filtering of contaminants is a critically important step to increase the accuracy of multivariate modeling (259).

Removing interferents prior to statistical analysis has numerous advantages. The filtering process allows for a more comprehensive annotation, and if artefacts are not removed, the relationships between samples may be distorted, potentially leading to a different biological interpretation (249, 256, 259). Typically a two-step process is

required to remove random analytical noise. First, chromatograms must be visually inspected to identify the signal intensity of the baseline. Following baseline signal assessment, any signals at or below the assigned baseline cutoff are then subtracted from the dataset. The remaining peaks should represent compounds associated with true chemical signals, although interpreting whether these signals originate from the sample or from background contamination remains a challenge (244).

Numerous types of chemical interferents can confound statistical analysis. Many interfering species are introduced as part of the sample preparation process itself, and may include solvent contaminants or polymeric interferents originating from sample vials, pipette tips or filter membranes. These contaminants will be consistent across samples, and are not the focus of this study. Some chemical interferents are not incorporated into the sample during sample preparation, but are introduced during the sample analysis step. These interferents originate from the analytical instrumentation, including silica capillaries, tubing, and HPLC column packing materials used for chromatographic separation (260, 261). We predict that chemical signals from these types of interferents may vary from sample to sample, and, consequently, will not be removable by blank subtraction (262).

Numerous approaches exist for identifying peaks exclusively associated with sample. One approach utilizes isotope-enriched nutrients in the growing media for mammalian cell culture (256, 263), plant tissue culture (264), or fungal culture (265). This approach produces labeling patterns that enable identification of which compounds are associated with the organism, and can highlight systematic changes in metabolic

processes due to various factors including environmental stress, genetic mutations, and disease state (256). Isotope labeling is an undeniably useful tool, but is only appropriate for applications assessing organisms grown in controlled environments.

Hierarchical cluster analysis (HCA) is a tool that uses an algorithm to produce a dendrogram that assembles variables or objects into a single tree, allowing users to visualize the similarity of the samples under analysis (244, 266). The HCA approach is usually used as a clustering tool to evaluate intra- and inter-group similarities and differences, similar to principal component analysis (PCA) (244, 267). In one study, HCA was used as a filtering tool to identify fragment ions associated with contaminant peaks (268). This process was used to overcome a common problem in gas chromatography-mass spectrometry GC-MS analysis, which is the production of molecular fragments originating from background contaminants such as fiber material. Using HCA, the investigators clustered the fragments in their samples, and identified and subsequently removed a cluster of masses associated with non-sample molecular fragments (268).

Here we present a new application of HCA, that of identifying chemical interferents in LC-MS analyses. We have chosen to use HCA over other clustering tools due to its distinct advantages for this application. First, HCA is a quantitative method to assess chemical similarity of different samples under analysis and the visualization in terms of a dendrogram makes it easy to assess if the removal of interferents has been successful. In other approaches, including K-Means clustering, density-based special clustering of applications with noise (DBSCAN), and PCA, the similarity of individual

points is often difficult to determine by eye and dependent on the components being graphically displayed. Furthermore, supervised analyses such as partial least squares-discriminant analysis (PLS-DA) require dependent variables that are able to separate contaminating interferents from discriminating compounds originating in the samples, and are not possible for this application. The goal of this study is to identify interferents that are introduced to the sample during the analysis process. This is achieved by comparing triplicate injections of the same sample (technical replicates) using HCA. Because the sample composition across replicate injections is identical, it is our expectation that chemical entities that vary across replicates will be interferents originating from analytical instrumentation, and that their removal will improve the quality of the data.

Experimental Section

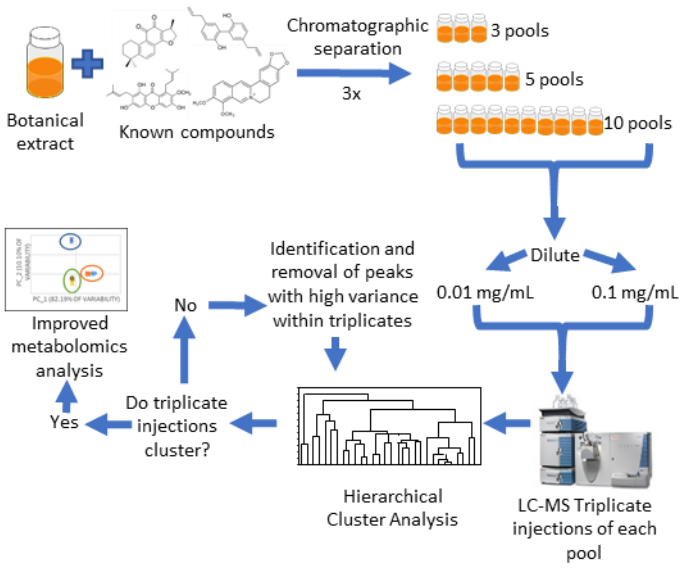
Sample preparation

The sample used for this study was produced as part of a separate project with the goal of optimizing the workflow for chemometric analysis in natural products research. These same samples were selected as a basis for the current study because they provided a good test case for evaluating chemical interference. To prepare the samples, a simplified extract of the botanical *Angelica keiskei* Koidzumi was spiked with four known compounds: alpha-mangostin (1% of total extract mass), cryptotanshinone (2% of total extract mass), magnolol (7% of total extract mass), and berberine (15% of total extract mass). Details about the method of extract preparation and can be found in Appendix A (Protocol S1).

Fractionation procedure

The spiked *Angelica keiskei* root extract was divided into three equal portions and subjected to reversed-phase HPLC separations. All three separations were conducted using the same gradient using a Gemini-NX reversed phase preparative HPLC column (5µm C18, 250 × 21.20 mm; Phenomenex, Torrance, CA, USA) at a flow rate of 21.4 mL min^{-1}. Starting conditions were 30:70 ACN:H_2O, which was increased to 55:45 over 8 min. Over the next two min., conditions were increased to 75:25, after which they were increased to 100% ACN by 28 min. Finally, the solvent composition was held at 100% ACN for two min.

Chromatographic separation was completed three times, with each separation yielding 90 test tubes. To evaluate the effect of sample complexity on hierarchical clustering analysis and data filtering approaches, we varied the number of pools in which the resulting tubes were combined. A "pool" is defined as a set of chromatographic fractions (in this case, multiple individual test tubes) that are combined together following chromatographic separation. The first set of 90 tubes was combined into three pools containing 30 tubes each, representing our most chemically complex samples. The second set of 90 tubes was combined into five pools consisting of 18 tubes each, and the final set of 90 tubes was combined into ten pools, each containing 9 tubes (**Scheme 1**). Each pool was dried under nitrogen and resuspended prior to LC-MS analysis.



Scheme 1. Workflow for Subset Preparation and Subsequent Analysis. A botanical mixture spiked with the known compounds berberine, magnolol, cryptotanshinone, and alpha-mangostin was fractionated three times and separated into equal sample sets containing 3, 5, or 10 final pools. The resulting pools were suspended at 0.1 or 0.01 mg mL^{-1} (reported as mass of dry extract per volume solvent) in methanol for UPLC-MS analysis. Each data subset was analyzed using hierarchical cluster analysis (HCA) before and after filtering to remove chemical interferents.

Mass spectral analysis

Full scan ultraperformance liquid chromatography-mass spectrometry (UPLC-MS) analysis was conducted on each pool using a Thermo-Fisher Q-Exactive Plus Orbitrap mass spectrometer (ThermoFisher Scientific, MA, USA) connected to an Acquity UPLC system (Waters, Milford, MA, USA) with reversed phase UPLC column (BEH C18, 1.7 µm, 2.1 x 50 mm, Waters Corporation, Milford, MA, USA). All pools were analyzed in triplicate at two different concentrations (0.1 mg mL^{-1} and 0.01 mg mL^{-1} in methanol, where concentration is expressed as mass of pool per volume of solvent), with 3 µL injections. The gradient was comprised of solvent A (water with 0.1% formic acid) and solvent B (acetonitrile with 0.1% formic acid). The gradient began with 90:10

(A:B) from 0-0.5 min, and increased to 0:100 (A:B) from 0.5-8.0 min. The gradient was held at 100% B for 0.5 min, before returning to starting conditions over 0.5 min and held from 9.0-10.0 min. Mass analysis was performed separately in both positive and negative ion modes over a *m/z* range of 150-1500 with the following settings: capillary voltage at -0.7 V, capillary temperature at 310°C, S-lens RF level at 80.00, spray voltage at 3.7 kV, sheath gas flow at 50.15, and auxiliary gas flow at 15.16. The top four most intense ions were fragmented with CID of 35.0.

Baseline correction and hierarchical cluster analysis

Baseline correction/MZmine parameters

UPLC-MS data collected in both negative and positive modes were individually analyzed, aligned, and filtered utilizing MZmine 2.21.2 software (http://mzmine.sourceforge.net/) (239). Raw mass spectral data from triplicate injections of each pool within the three sets were uploaded for peak picking into MZmine. Chromatograms were constructed for all *m/z* values with peak widths greater than 0.1 minute, after which they were simplified using algorithms applied to recognize individual peaks. The peak detection parameters were set as follows: noise level (absolute value) at 2.0×10^6 (positive mode, 0.1 mg mL^{-1} samples), 1.0×10^7 (positive mode, 0.01 mg mL^{-1} samples), and 1.0×10^6 (negative mode, both 0.1 mg mL^{-1} and 0.01 mg mL^{-1} samples), *m/z* tolerance of 0.0001 Da or 5 ppm, and an intensity variation tolerance at 20%. Peaks were aligned if their masses were within 5 ppm and their retention times differed by less than 0.2 min from one another. Peak list filtering and retention time alignment were completed to produce an aligned peak list. The resulting data matrix, consisting of *m/z*,

retention time, and peak area, was imported into Excel (Microsoft, Redmond, WA, USA). Peak lists for positive and negative ions were combined, and separate data sets were generated for high and low concentration samples. No further pre-processing of data sets was completed before hierarchical cluster analysis and data filtering.

Hierarchical cluster analysis and chromatograph visualization

Hierarchical clustering analysis and resulting filtering protocols were completed using Sirius version 10.0 statistical software (Pattern Recognition Systems AS, Bergen, Norway) (240, 269). For this analysis, an average-linkage algorithm (270) was used to cluster objects. Euclidean distance was used as a metric to evaluate object similarity.

Six data sets were produced (three high-concentration and three low-concentration data sets containing 3-, 5-, or 10-pools and their triplicate injections) and inspected using HCA. A dataset was considered correctly clustered only when all triplicate injections were linked to each other before being linked to other samples in the dendrogram. If triplicate injections did not cluster, spectral variables (mass/retention time pairs) were inspected for each set of triplicates. Since highly abundant or highly ionizable compounds inherently have higher count variance, the contaminant masses were identified by examining relative variance within each set of technical replicates, defined by equation 1**.** Variance (s_k^2) represents the sum of the squared differences of each compound's peak area (x_k) from the mean of its peak area within replicate injections ($\bar{x}_k$), divided by the number of replicates (N_r). Relative variance of peak *k* in sample *i* ($RV_{k,i}$) was calculated by dividing the variance within replicates by the mean.

$$RV_{k,i} = s_k^2/\bar{x}_k, \text{ where } s_k^2 = \Sigma(x_k - \bar{x}_k)^2 / N_r \qquad \text{(equation 1)}$$

Of course, it is possible that a non-interferent peak may show high variability in peak area from injection to injection, particularly if it co-elutes with another sample component that impacts its ionization. To minimize the risk of removing false positives, we chose to sort variables from high to low RV based on their average relative variance values ($\overline{RV}$), defined by equation 2. Even if in one sample the ionization of a given sample component was affected by matrix effects, it is unlikely that this response would be consistent across samples with different chemical constituents. Average relative variance for the peak k ($\overline{RV}_k$) was calculated by dividing the sum of the relative variances (calculated *within* each pool's set of replicate injections: $RV_{k,1}$, $RV_{k,2}$, . . . $RV_{k,p}$) by the number of pools (N_p).

$$\overline{RV}_k = (RV_{k,1} + RV_{k,2} + \ldots + RV_{k,p}) / N_p \qquad \text{(equation 2)}$$

Using equation 2, the variables with the highest average relative variance were identified and removed, and intermittent hierarchical cluster analysis was conducted. To assist the analysis, spectral variable plots were utilized to visualize mass/retention time pairs identified using the selected relative variance cutoff as contaminants as well as their corresponding ions. The ions that demonstrated peak area variability within triplicate sets higher than the selected threshold, as well as their associated isotopes, in-source clusters, or fragments were removed from analysis. HCA was repeated to visualize how well samples clustered once the contaminants were removed. This was repeated until triplicate injections of each sample were linked before being linked to other samples in all datasets.

Results and Discussion

Hierarchical cluster analysis and data filtering

The goal of this study was to identify and remove chemical contaminants from mass spectral data sets. Towards this goal, HCA was conducted on six sets of pools, where each pool was injected in triplicate (technical replicates) into the UPLC-MS system. Each dataset was analyzed by HCA after baseline correction and peak alignment(244). It was expected that the replicates would show high chemical similarity and cluster together in the dendrogram. Before filtering out chemical interferents with high peak area variability within technical replicates, however, triplicate injections clustered together in only one of the six data sets—the three-pool subset analyzed at 0.1 mg mL^{-1} (Table 8, Figure 19).

Table 8. Summary of Hierarchical Clustering Analysis Before and After Data Filtering.

Sample Set	Percentage of Correct Triplicate Clusters Before & After Filtering Analysis (Before, After)	Average Dissimilarity Score* Before & After Filtering Analysis (Before, After)
Three pool set, 0.1 mg mL^{-1}	100%, 100%	5.23×10^9, 3.23×10^9
Three pool set, 0.01 mg mL^{-1}	33%, 100%	6.17×10^9, 8.62×10^8
Five pool set, 0.1 mg mL^{-1}	60%, 100%	6.18×10^9, 2.34×10^9
Five pool set, 0.01 mg mL^{-1}	20%, 100%	5.71×10^9, 4.54×10^8
Ten pool set, 0.1 mg mL^{-1}	40%, 100%	3.05×10^9, 1.36×10^9
Ten pool set, 0.01 mg mL^{-1}	0%, 100%	8.13×10^9, 3.72×10^8

*Average dissimilarity scores were computed in Sirius 10.0 (240, 269) and represent n-dimensional Euclidean distance values.

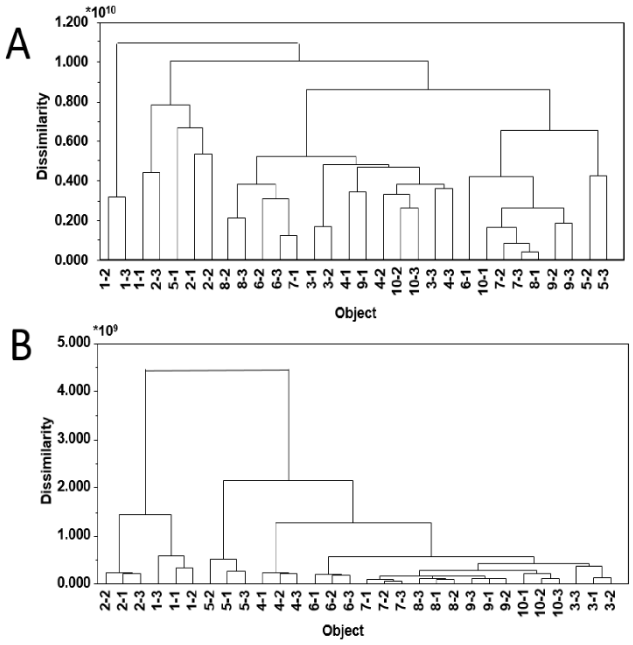


Figure 19. Euclidean Dendrograms of the Ten-Pool, 0.01 mg mL^{-1} Data Subset Before (A) and After (B) Filtering Analysis. Samples have been identified first by their pool number followed by the injection number. For example, 1-1 is the first pool, and first injection of three technical replicates.

On inspecting the data sets, it was determined that certain masses were present in all samples but did not display consistent peak area across triplicates. We hypothesized that these masses were chemical interferents and not truly sample components. Thus, removing these masses from the data sets should result in the expected clustering of replicates. To identify the variables representing chemical interference, the relative variance of each variable was calculated for each set of triplicate injections as defined in equations 1 and 2. The relative variance cutoff was determined by reducing the threshold until dendrograms showed the expected classification of replicate injections. The dataset was considered filtered when replicates clustered together before clustering to additional samples. Contaminant peaks were assigned as those that had an average relative variance

ratio (across all pools) greater than 1.0×10^7 for low concentration data sets, and 4.1×10^7 for high concentration data sets. The same interferents were identified in both subsets, though more interferents were identified using the low concentration data subsets.

In addition, each chromatogram was visually inspected using a spectral variable plot, in which the mass/retention time of each unique spectral variable was plotted on the x-axis, and corresponding peak area of that variable was plotted on the y-axis (Figure 20A). Ions that were part of the sample, including the known compounds spiked into the mixture, showed consistent peak area across triplicate injections (Figure 20B), whereas purported contaminants typically did not (Figure 20C). Spectral variable visualization also enabled the identification of peaks associated with the contaminant masses, such as ^{13}C isotopes and in source clusters and fragments, which were not identified based solely on the mathematical approach. For example, two mass/retention time pairs were identified at *m/z* 744.201 and 744.211 using the relative variance cutoff. Upon spectral variable inspection, additional isotope peaks and mass spectral artefacts associated with this contaminant were identified (e.g. *m/z* 746.188, 746.198, and 746.208, Appendix B, Table S1), despite their low relative variance (Figure 20C). Removing these ions improved clustering, allowing for a more complete representation of contaminants. Background contaminants with high peak area variation between triplicate injections, as well as their associated masses, were removed from the peak list, and HCA was repeated. Following the removal of these compounds, triplicates clustered in all six sample subsets and the average dissimilarity score of technical replicates decreased (Table 8). An example dendrogram before and after filtering is shown in Figure 19. It is important to

note that there is the possibility that true sample components may share the same *m/z* and retention times as isotopes and mass spectral artefacts and be accidentally removed during this process. In some cases, fragmentation patterns can be evaluated to assess whether or not these masses are truly associated with contaminant peaks showing high relative variance. This may not always be possible, however, so users familiar with the analytical instrumentation and the biological sample under analysis should conduct this part of the filtering process carefully, with the goals of the project in mind.

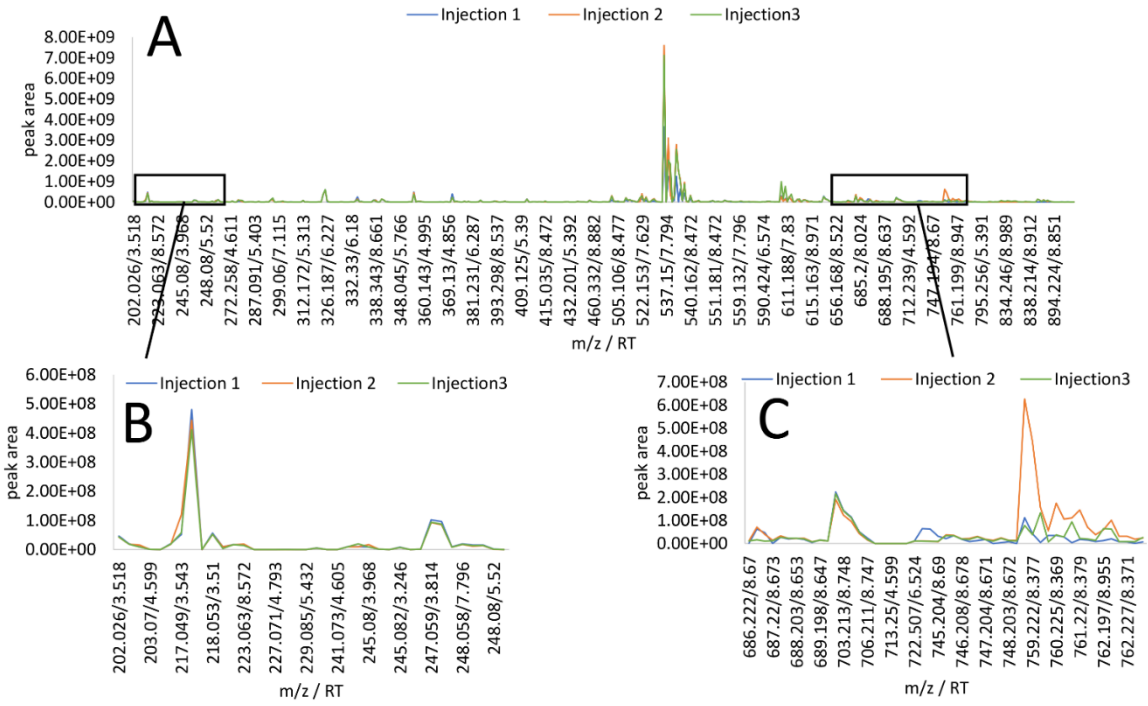


Figure 20. Spectral Variable Inspection of Triplicate Injections from the Second Pool from the Five-Pool, 0.01 mg mL^{-1} Data Subset. 20A. Overlaid spectral variable plots of triplicate injections in which peak areas of each variable are plotted for comparison. 20B. Spectral variables associated with the sample under analysis. Overlapping traces are consistent from injection to injection. 20C. Spectral region associated with chemical contamination showing a variance/mean peak area ratio greater than 1.0×10^7.

Sources of Contamination

Source of chemical interferents with high peak area variability

As analytical instruments have become more sensitive and more high-throughput, the list of potential interferents detected grows (271). During chromatographic separation and mass spectral analysis, the sample comes into contact with a variety of surfaces that could lead to chemical contamination not associated with the sample, such as polymeric interferences from plasticware and tubing (271). We hypothesized that ions demonstrating high peak area variation between triplicate injections were due to chemical interferents coming from sources such as these. These contaminants (Appendix B, Table S1) were consistent in their identity (although not peak area) across data sets. Of the 128 contaminant peaks removed from analysis, 22 were tentatively identified (using accurate mass data) as associated with polysiloxanes as reported by Keller et al. (271). Indeed, polysiloxanes are found in silica capillary tubes such as those used for UPLC-MS analysis as well as in column packing materials (260, 261).

To investigate our hypothesis that these contaminants originated from the analytical instrumentation, the accurate masses and retention times of common interferents were compared to blank injections containing methanol with no sample. Methanol blanks were included throughout the run. Of the 128 contaminants identified, 121 were present in at least one of the blanks. Interestingly, 44 of the interferent features were not found in every blank. Thus, it appears that the interferents originate from the UPLC-MS system itself, and not from the solvent alone (although it is possible that both the solvent and the UPLC-MS system might contain some of the same contaminants).

It is common practice in metabolomics analysis to subtract peaks contained within the blank from the data sets under study (262). However, our results (Figure 20) show that ion abundance of chemical contaminants can vary from injection to injection, so the list of contaminants removed will likely not be comprehensive using a simple blank subtraction. Indeed, when we produced a dendrogram of the 0.01 mg mL^{-1}, ten-pool set after subtracting peaks contained within one of the blanks, the triplicate injections only clustered together for two out of ten pools (Figure 21). Additionally, in cases where carryover occurs between sample and blank injections, it is possible that ions contained in the sample can be inadvertently removed by blank subtraction. The method proposed here in which replicate injections are compared to identify potential contaminants circumvents the problems associated with subtracting the peaks from a single blank injection.

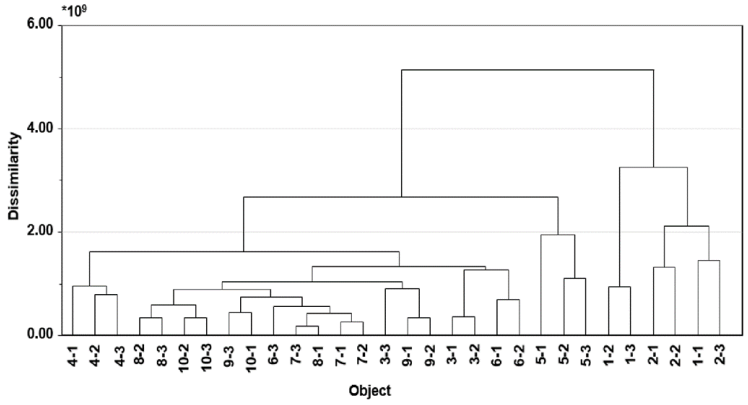


Figure 21. Euclidean Dendrogram of the Ten-Pool, 0.01 mg mL^{-1} Data Subset Following Subtraction of Masses Contained in One Blank from Analysis. This example illustrates that blank subtraction was insufficient since replicates do not cluster correctly.

The potential for false positives

We have illustrated that an important type of chemical interference originates from the LC-MS equipment itself, and have developed a method to minimize its contribution to metabolomics datasets. However, there is the potential for this approach to remove actual sample components that show high variability among replicate injections. Peak area variance can occur for a number of reasons, including matrix ionization effects, injection errors, and sample carryover from previous injections (272). Matrix effects leading to changes in compound ionization efficiency or mobility can result from interactions with other components of the sample. If a particular compound co-elutes with another sample component that impacts its ionization, for example, it may not show consistent peak area from injection to injection and could be identified as a false positive. Similarly, injection errors, in which the actual sample volume analyzed via LC-MS is different than expected, can lead to large differences in peak area across injections, even for true sample components (272).

Although there is a risk for removing false positives with the method proposed herein, the use of average relative variance as a metric to define contaminants (equation 2) reduces this risk. It is likely, for example, that in at least one set of triplicate injections a real sample component may be affected by matrix effects and consequently show a high relative variance. It is unlikely, however, for this matrix effect to be consistent across all samples under analysis, and the high relative variance value from one sample should be normalized by averaging with low relative variance values from other samples. The same is true for injection errors and sample carryover.

It is of course possible that even when using average relative variance, we may unintentionally remove important sample components from our datasets. To reduce this risk further, we recommend the use of internal standards. These internal standards should consist of compounds possessing diverse properties and should not be found in the biological sample under analysis (272). Differences in peak areas of these internal standards can allow researchers to identify samples that may have been compromised by sample injection errors or matrix effects, and back-calculations can be used to correct for changes resulting from these processes.

Complementary quality control practices to improve metabolomics datasets

It is important to note that the types of chemical contaminants identified using the approach presented here are only those contaminants that vary distinctly from replicate injection to injection. Interferents that originate from the sample preparation process will be consistent across technical replicates and not identified with the HCA approach demonstrated here. Complex calibrants such as process blanks, which do not contain biological material but have undergone the same chemical treatment as biological samples (272) should be included in LC-MS analyses to identify interferents resulting from sample preparation. Compounds found in process blanks may represent some of the same contaminants identified using this HCA approach (if the sample had gone through some sort of chromatographic separation step before LC-MS analysis, for example), but will likely contain additional chemical contaminants including pipette tip contaminants and extraction solvent impurities (272, 273). Although we have illustrated that the inclusion of solvent blanks is not sufficient to remove all contaminants from analysis,

blank runs are still undeniably important, as they allow researchers to define an appropriate baseline cutoff, estimate background noise, and evaluate carryover effects (262, 272).

Effects of sample number and concentration on dendrogram analysis

To evaluate the effect of sample number and concentration on filtering analysis, we compared sets containing three-, five-, or ten-pools at concentrations of 0.1 mg mL^{-1} and 0.01 mg mL^{-1} (expressed as mass of the mixture per volume solvent). Because the three-, five-, and ten-pool subsets all originated from the same starting mixture, the resulting pools will be the most complex with the lowest number of pools (Scheme 1).

Data sets containing greater numbers of samples were more impacted by chemical interferents, as were samples injected at the lower concentration of 0.01 mg mL^{-1} (Table 9). For example, the average dissimilarity scores, calculated by averaging scores from the 0.1 mg mL^{-1} subsets and the 0.01 mg mL^{-1} subsets, respectively, were higher in low-concentration groups when compared to their high-concentration counterparts (6.67×10^9 versus 4.82×10^9, respectively). Following filtering analysis, high-concentration groups showed greater dissimilarity scores than the low- concentration subsets (2.71×10^9 and 8.39×10^8, respectively). After filtering, both high- and low-concentration subsets displayed lower dissimilarity scores than they did preceding data filtering, indicating that the contaminant peaks contributed to the high dissimilarity between triplicate injections.

The results of these comparisons are illuminating, and suggest that metabolomics studies of simpler samples may be more impacted by chemical interferents. Indeed, the three-pool dataset, regardless of injection concentration, consistently showed the highest

number of correct clusters before filtering analysis. Because they contain more compounds that are consistent between triplicate injections, the varying concentrations of contaminants have less effect on the overall clustering of more complex pools. With the simpler pools in the ten-pool dataset, the effect of high variability in peak area of contaminant peaks has a greater influence on the overall model. Similarly, the effect of contaminant interference appears to be greater with low-concentration injections, presumably because the contaminant peaks have larger relative peak areas in these subsets. This is an important point, because metabolomics analysis is often focused on identifying very low-abundant peaks from highly complicated samples. As such, filtering analysis to remove interferents may be critical for success.

PCA scores and loadings

Principal Component Analysis (PCA) is one of the most commonly employed tools in metabolomics data analysis and is used to group objects by chemical similarity (132). Groupings of objects can be visualized in a PCA scores plot, and the variables contributing to the groupings can be assessed using a PCA loadings plot. PCA was used here as an alternative technique to HCA to assess the similarity of triplicate injections.

As an example, the ten-pool, low-concentration dataset was subjected to PCA before and after removal of the chemical interference ions. Before filtering, untargeted metabolomics analysis of these pools yielded 467 marker ions with unique retention time-*m/z* pairs. The resulting PCA model comparing the pools was comprised of two components explaining 81.8% of the variance (component 1: 53.1%, component 2:

28.7%). The technical replicates of each pool did not cluster on the resulting scores plot, indicating that interferents have a severe impact on clustering analysis (Figure 22A).

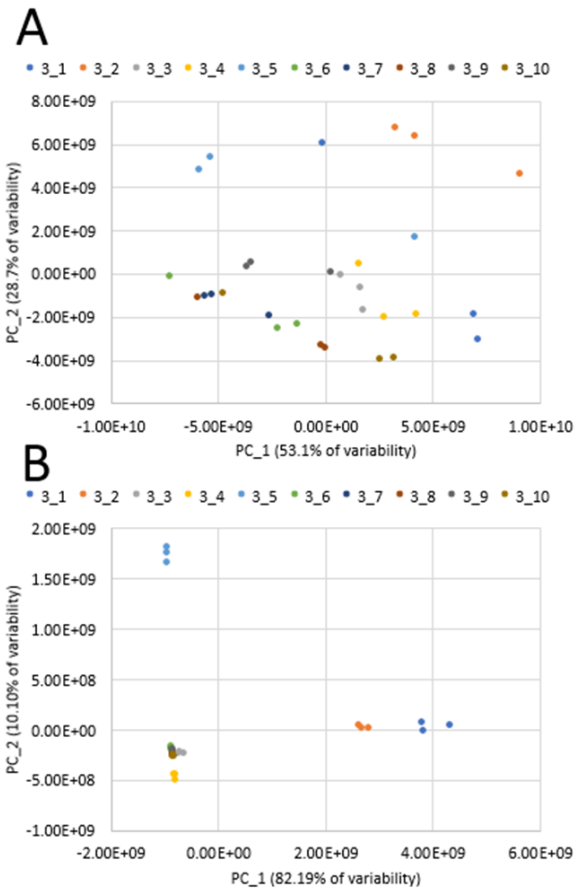


Figure 22. PCA Scores Plots Before (A) and After (B) Data Filtering of the Ten-Pool, 0.01 mg mL^{-1} Data Subset. 22A. Technical replicates are not overlaid on the plot, and clustering of groups is difficult to visualize. 22B. Technical replicates are overlaid as expected, and there is distinct separation between groups.

Following spectral variable inspection and removal of contaminant masses, a new PCA model was produced, this time containing 339 ions. The two-component model explained 92.29% of the variance (component 1: 82.19%, component 2: 10.10%). With this model, triplicate injections are overlaid on the plot, indicating that statistical analysis was virtually unaffected by chemical interferents (Figure 22B). If chemical interferents

that varied from injection to injection were still impacting the analysis, we would expect that triplicates would not cluster in the scores plot, as evidenced with Figure 22A. Any contaminants that remain in the dataset are likely consistent in peak area across all samples under analysis, and will consequently have little to no effect on the resulting principal component analysis. Additionally, clusters between groups of pools are more distinct following contaminant removal, ultimately improving both repeatability and cluster identification.

The PCA loadings plot before analysis is revealing (Figure 23) and shows that many of the loadings resulting in separation of pools are associated with interferents.

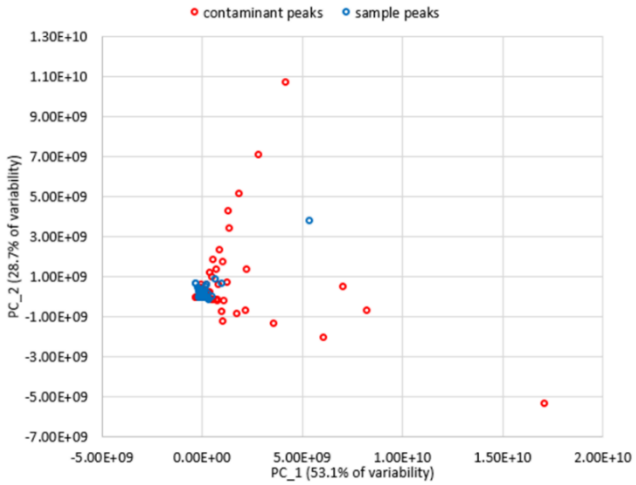


Figure 23. PCA Loadings Plot of the Ten-Pool, 0.01 mg mL^{-1} Data Subset Before Filtering of Chemical Interferents. Most of the variables contributing to group separation are contaminant peaks.

Hypothetically, it would be possible to utilize PCA loadings plots of triplicate injections to visualize which compounds contribute to separation of chemical replicates (Figure 24). Because this loadings plot is comprised *solely of a set of triplicate injections*, all variables should be clustered. However, this is clearly not the case, and any variables that lead to

group separation are due to chemical interference introduced *after* sample injection. From Figure 24B, it is apparent that contaminant variables are responsible for separation between triplicate injections. Contaminant and sample variables do begin to overlap in the center of the plot, making visual interpretation challenging without knowledge of mixture components. The loadings plot of one set of triplicate injections (first pool of the ten-pool set, 0.01 mg mL^{-1}) is difficult to interpret, and contaminant peaks can only be arbitrarily identified (Figure 24). This was a common problem across all datasets.

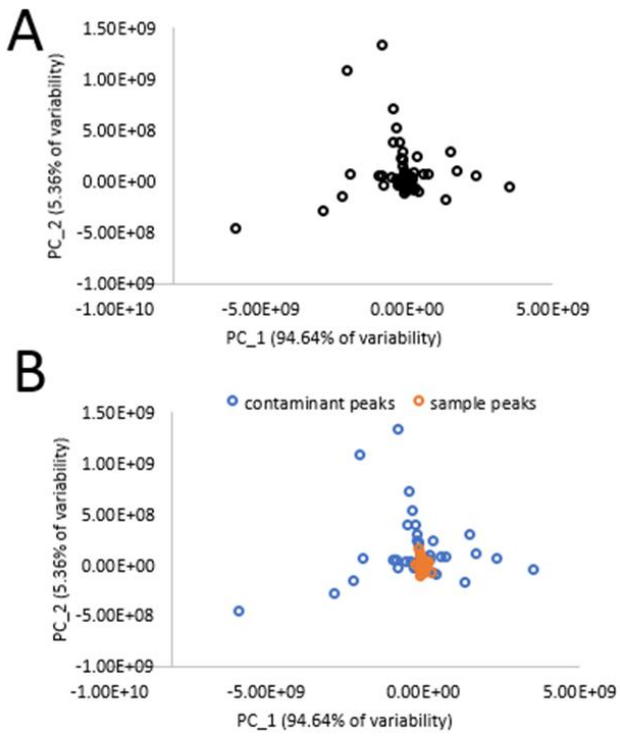


Figure 24. PCA Loadings Plot of Triplicate Technical Replicates from Pool One of the Ten-Pool, 0.01 mg mL^{-1} Data Subset. 24A. Loadings plot illustrating all variables contributing to group separation. 24B. Color-coded loadings plot allowing visualization of contaminant and sample peak influence on group separation. Many of the chemical contaminants are close to the center cluster and would not be reliably identified using PCA loadings alone.

Conclusions

Robust data pretreatment is necessary to extract reliable information from mass spectrometry data sets. The results presented here demonstrate that HCA of technical replicates is a valuable tool for data pretreatment by enabling the identification and removal of certain interferents. In its current form, however, there is still a considerable amount of user-intervention required. Further developments should focus on automating this approach as much as possible so that users do not have to iteratively filter their data by hand and define the relative variance cutoff. However, identifying peaks that are associated with interferents yet do not show high relative peak area variance will still require identification by the user, and this application will vary from experiment to experiment and depend on the goals of the project itself.

It is often assumed that chemical interference in mass spectral data is consistent across samples, and should, therefore, be removable by blank subtraction. On the contrary, here we show that certain chemical interferents can vary in signal intensity across technical replicates. Such interferents can be identified and removed with the approach presented here. It is particularly important to identify and remove these types of interferents, given that the very premise of metabolomics experiments is that the compounds that vary among samples are likely to be chemically or biologically relevant. Many studies are conducted without technical replicates, and the results of the current study show the potential limitation of such an experimental design and demonstrate a straight-forward alternative strategy.

Acknowledgements

The authors would like to thank Strictly Medicinal Seeds® for their provision of plant material used in this project. Research reported in this publication was supported in part by the National Center for Complementary and Integrative Health of the National Institutes of Health under award numbers 5 T32 AT008938 (fellowship to LKC), U54 AT008909 (NaPDI, Center of Excellence for Natural Product Drug Interaction Research), and R01 AT006860.

CHAPTER V

OPPORTUNITIES AND LIMITATIONS FOR UNTARGETED MASS SPECTROMETRY METABOLOMICS TO IDENTIFY BIOLOGICALLY ACTIVE CONSTITUENTS FROM COMPLEX NATURAL PRODUCT MIXTURES

Chapter V is formatted for and published in Journal of Natural Products. Caesar, L.K., Kellogg, J.J., Kvalheim, O.M., Cech, N.B. J Nat Prod. 2019, *in press*.

Caesar, L.K. conceived of the idea for this project, collected and processed all data, and wrote the manuscript. Kellogg, J.J. provided input for biochemometric analysis and assisted with experimental design. Kvalheim, O.M. created the software used for this project (Sirius) and offered technical advice for statistical analysis. Cech, N.B. assisted in the development of the research project and provided edits and suggestions throughout manuscript preparation.

Introduction

Untargeted metabolomics is poised to make an impact in many areas of research, including studies to understand disease pathogenesis (247), to assess food quality and authenticity (250), to monitor the environmental quality of water resources (274), for biomarker identification (251-253), and for drug discovery (18, 132, 248, 254). Mass spectrometry is a leading tool for generation of untargeted metabolomics datasets, largely due to the applicability of this technique to provide quantitative and qualitative data on many metabolites simultaneously across a wide range of concentrations (244). Mass

spectrometry metabolomics yields high-dimensional datasets that offer a detailed chemical picture of the organism in question. These data can be employed in a discovery-driven approach to guide understanding of complex mixtures and enable linkage between a biological effects and the chemical profile of a given organism (8, 275). However, the interpretation of mass spectrometry metabolomics datasets is complex, requires multivariate data analysis methods, and may be confounded by experimental artefacts (276, 277). There is currently lack of consistency in the field regarding methods for collecting and interpreting metabolomics datasets, and concerns have been raised as to the reproducibly of conclusions drawn from metabolomics datasets (256). In light of these concerns, the work described herein was undertaken to rigorously evaluate the advantages and limitations of metabolomics approaches for one specific application – that of identifying biologically active compounds in complex natural product extracts.

Natural products such as plants, fungi, marine organisms, and bacteria have been utilized as medicines throughout history and continue to provide lead compounds effective against human diseases (222, 223). However, due to the diversity of identity and abundance of compounds produced by natural products, it remains challenging to assign bioactivity to individual components in such mixtures . The traditional solution to this problem is bioassay-guided fractionation (220, 221), in which active extracts and subsequent fractions are subjected to iterative chromatographic separations and biological evaluation until individual active compounds have been isolated. This process, despite its historical contribution to the discovery of important medicinal compounds, tends to be

biased towards the most abundant, easily detectable, and/or easily isolatable compounds in a given mixture (131, 220). To overcome abundance bias, trace constituents can be isolated, but it is impractical to isolate all trace compounds given that natural products often contain hundreds or even thousands of constituents (13). In recent years, multiple different groups have sought to guide active constituent identification by integrating metabolomics data (chemical profiles) with biological activity data (biological activity profiles), enabling isolation efforts to be targeted towards active rather than abundant constituents (132, 149, 150, 278). Approaches that employ multivariate statistics to interpret combined chemical and biological datasets are broadly referred to as "biochemometrics."

Several different data analytical approaches are used as tools in biochemometrics analyses. Due to the large number of variables compared to the number of samples analyzed, data from complex mixtures possess a high degree of collinearity. This poses a problem for ordinary multiple regression models, but partial least-squares (PLS) regression is capable of integrating aspects from both multiple regression and principal component analysis, making it a good starting point for biochemometrics analysis (148). The resulting multicomponent PLS models are, however, often challenging to interpret. Several strategies have been developed for deciphering the meaning of PLS datasets (132, 149, 150, 278). Two graphical representations, the S-plot and the selectivity ratio plot, can be employed to visualize the information in PLS models and determine which components are likely to contribute to an observed biological activity.

The S-plot provides an avenue for identifying predictive components by plotting covariance and correlation of loading variables. Using an S-plot, constituents that have both high covariance and high correlation with the dependent variable in question can be identified (132, 279). S-plots have been successfully used in many studies to identify potential biomarkers for disease treatment (280), to authenticate the origin of food crops (281), and to identify medicinal compounds from botanical sources (282, 283), among others. However, the criterion of high covariance favors the identification of abundant compounds, while trace bioactive constituents may go undetected (233). Identifying points of interest can also become challenging due to the large number of spectral variables (132). The selectivity ratio plot (269) overcomes the abundance bias inherent to the S-plot by transforming the PLS components to enable quantification and ranking of each variables´ impact on the modelled response, i.e. bioactivity, independent of the abundance of the variables. The explained variance on the predictive PLS component is compared to the residual variance for each constituent to produce a selectivity ratio (269), which is a measure of the predictive contribution of each variable to bioactivity.

In a recent study, fungal extracts were subjected to biochemometric analysis to determine which constituents were responsible for biological activity (ability to inhibit bacterial growth) (132). Selectivity ratio plot analysis correctly identified altersetin from the fungus *Alternaria* sp. as the active constituent despite its low abundance without being confounded by false positive results. In a parallel study, both the S-plot and the selectivity ratio plot were successful in identifying the major component macrosphelide A as the active constituent from *Pyrenochaeta* sp. (132). A similar investigation was

undertaken to identify compounds that enhanced the antibacterial efficacy of the alkaloid berberine within the botanical medicine *Hydrastis canadensis* (18). Biological activity data were combined with untargeted metabolomics data to produce selectivity ratio plots, which successfully identified known synergistic flavonoids and a new compound, 3,3'-dihydroxy-5,7,4'-trimethoxy-6,8-*C*-dimethylflavone, which also possessed synergistic activity (18). This study illustrated the applicability of selectivity ratio analysis to predict active components of complex botanical mixtures. It was possible to identify false positive results because they did not possess activity following isolation. However, without isolating every trace constituent in the mixture, the biochemometric models were unable to identify the frequency of false *negative* results.

With the work described herein, we aimed to evaluate the occurrence of false positive and false negative results when biochemometric analysis is conducted using selectivity ratio plot analysis and to optimize experimental conditions and data processing approaches to minimize the occurrence of both types of false positives. Towards this goal, we generated mixtures containing an inactive botanical natural product extract spiked with known antimicrobial compounds (berberine, magnolol, cryptotanshinone, and alpha-mangostin, Figure 25. compounds **1-4**, respectively). Using these mixtures, we sought to assess the predictive power of selectivity ratio analysis combined with several data filtering and data transformation approaches for identifying active (antimicrobial) constituents based on chemical (metabolomics) and biological data.

Results and Discussion

Chromatographic separation and generation of simplified pools

A simplified extract of the botanical *Angelica keiskei* Koidzumi was spiked with four known constituents (compounds **1-4**) and split into three chemically identical samples. Each sample was subjected to the same reversed-phase chromatographic separation process with each run yielding 90 test tubes. These tubes were re-combined into three pools of 30 tubes (samples 1-1 through 1-3), five pools of 18 tubes (samples 2-1 through 2-5), or ten pools of 9 tubes (samples 3-1 through 3-10) to generate the simplified *A. keiskei* pools for biochemometric analysis and statistical comparison.

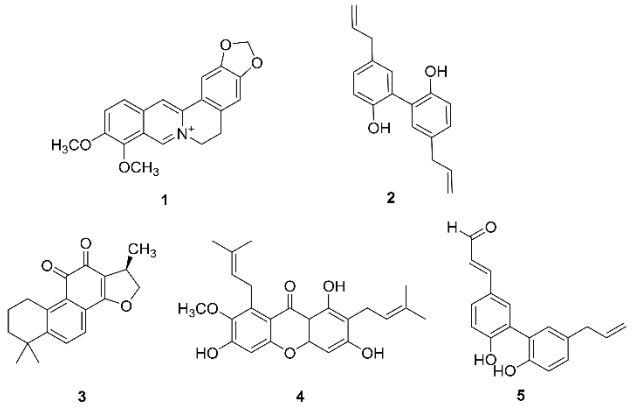


Figure 25. Bioactive Compounds Utilized in this Study.

Biological activity assessment and confirmation of active compounds

Antimicrobial activity assessment

At the highest concentration tested (100 µg/mL), seven of the spiked *A. keiskei* pools completely inhibited the growth of *Staphylococcus aureus.* At 50 µg/mL, only four pools inhibited more than 80% of bacterial growth. At 25 µg/mL, none of the treatments

resulted in more than 50% inhibition. The results of these assays are summarized in Figure 26**.** None of the pools showed any activity at concentrations lower than 25 µg/mL.

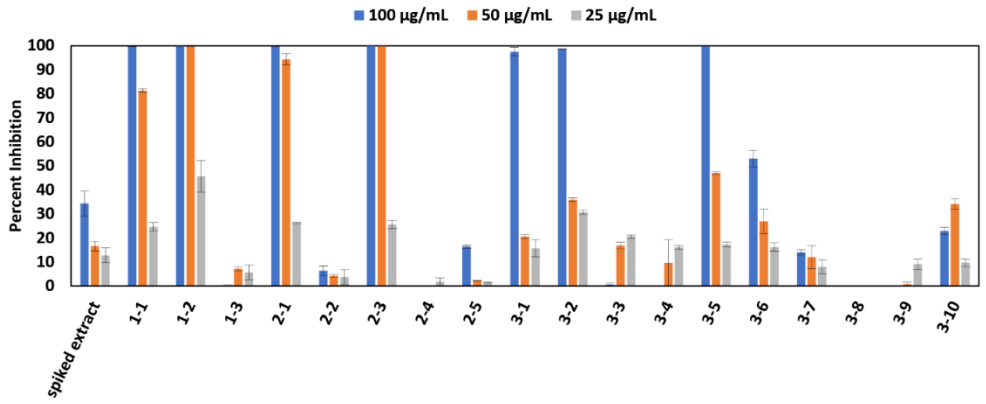


Figure 26. Antimicrobial Activity Data of the *A. keiskei* Root Extract Spiked with Known Antimicrobial Compounds (Spiked Extract) and Eighteen Chromatographically Separated Pools from the Original Spiked Extract. Pools labeled 1-1 through 1-3 represent samples resulting from chromatographic separation of the spiked *A. keiskei* root mixture into three pools, 2-1 through 2-5 represent samples from separation into five pools, and 3-1 through 3-10 represent samples from the ten-pool set. Growth inhibition of *Staphylococcus aureus* (SA1199) (238) is displayed as percent growth inhibition normalized to the vehicle control (broth containing bacteria but no antimicrobial compound) using OD_{600} values. Data presented are the results of triplicate analyses ± SEM. Pure compounds berberine (**1**), magnolol (**2**), cryptotanshinone (**3**), and alpha-mangostin (**4**) served as positive controls and their minimum inhibitory concentrations (75, 6.25, 12.5, and 1.56 µg/mL, respectively), are consistent with previous reports (132, 284-286).

Quantification of known compounds and predicted activity calculations

Concentrations of known active compounds berberine, magnolol, cryptotanshinone, and alpha-mangostin (compounds **1-4**) were quantified using external calibration curves (Appendix C, Figure S14). The dose response curves of pure compounds (Appendix C, Figure S15) were then used to predict their biological activity at 100 µg/mL. A comparison of this predicted total activity and the observed bioactivity of the relevant pool at 100 µg/mL is shown in Figure 27. Pools 1-1, 2-1, and 3-1 contained 50-75 µg/mL of berberine (compound **1**), which was predicted to result in 75-

100% growth inhibition. Magnolol (compound **2**) was predicted to inhibit bacterial growth in the spiked extract before fractionation, as well as in pools 1-2, 2-3, and 3-5.

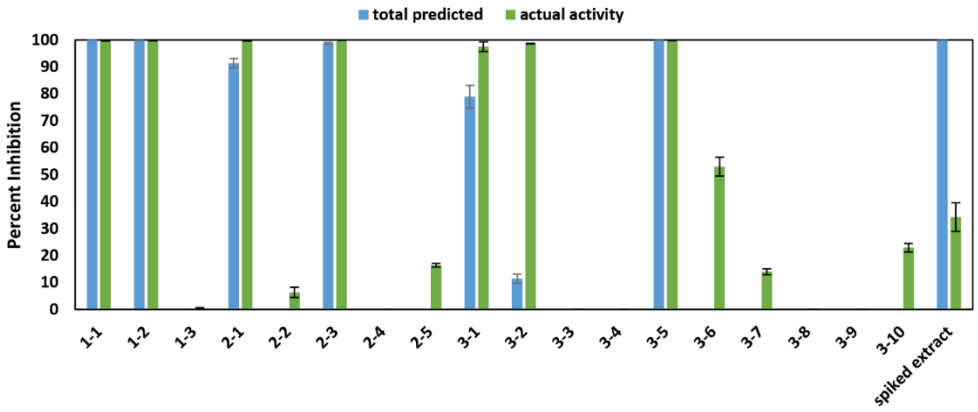


Figure 27. Predicted versus Actual Antimicrobial Activity of *A. keiskei* Spiked Extract and Pools at 100 µg/mL. Predicted antimicrobial activity was calculated by quantifying compounds **1-4** (berberine, magnolol, cryptotanshinone, and alpha-mangostin) in each pool and using these values to calculate predicted contribution to activity (via dose response curves). Actual activity values represent percent growth inhibition of *Staphylococcus aureus* (SA1199) (238) normalized to the vehicle control (broth containing bacteria but no antimicrobial compound) turbidimetric OD_{600} values. Data presented represent results of triplicate analyses ± SEM. Positive control data are the same as described for Figure 26.

These pools contained between 5 and 10 µg/mL of magnolol, contributing 85-100% to the predicted activity. Cryptotanshinone (compound **3**) was predicted to inhibit 15% of bacterial growth in pool 1-2 (containing approximately 3 µg/mL), and 40% of growth in the unseparated mixture (which contained approximately 5 µg/mL). Alpha-mangostin (compound **4**) was not present at concentrations relevant for biological activities in any of the pools tested.

The observed activity of six of the active pools (1-1, 1-2, 2-1, 2-3, 3-1, and 3-5) matched the predicted activity from the calculated concentration of a particular bioactive constituent in each of those pools; thus, the activity was explained almost completely by

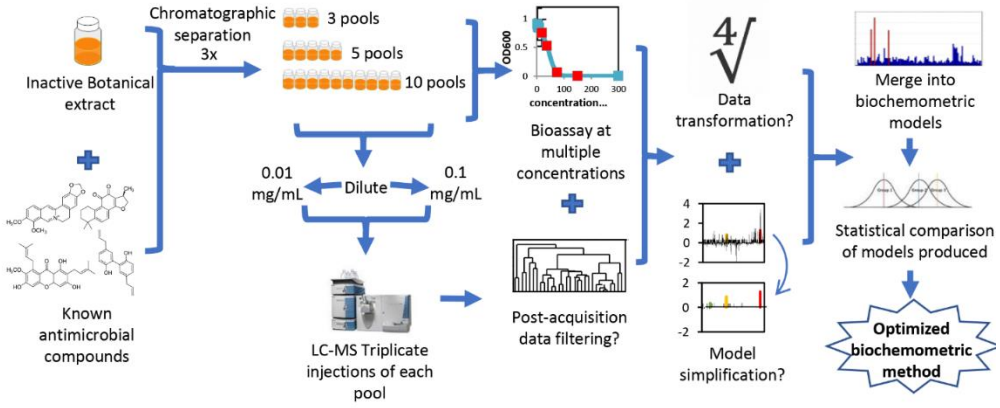
the predicted contributions of berberine and magnolol. Pools 3-2 and 3-6 demonstrated 100% and 50% activity, respectively, which could not be attributed to the predicted contributions of berberine and magnolol. Interestingly, the spiked *A. keiskei* mixture was predicted to completely inhibit bacterial growth, but only illustrated approximately 35% inhibition. This observation, which suggests antagonistic activity of the mixture, is discussed in detail later (see section: Assessment of Combination Effects in Spiked *A. keiskei* Mixture).

Selectivity ratio analysis and comparison of protocols

General findings

PLS models for predicting active compounds were produced and visualized using selectivity ratio analysis. With these selectivity ratio models, each ion detected (represented by a *m/z* retention time pair) is plotted on the x-axis and its corresponding selectivity ratio is shown on the y-axis. High selectivity ratio values represent ions that are most strongly associated with biological activity. We sought to produce eighteen different models utilizing samples from datasets with three different numbers of chromatographic pools (3, 5, or 10), bioactivity obtained at three different concentrations (25, 50, or 100 µg/mL), and profiles for two different pool concentrations injected into the LC-MS system (0.1 or 0.01 mg/mL). In each model, selectivity ratios were ranked from high to low, and the rankings of active compounds berberine and magnolol were evaluated. These compounds should have been identified as the top two contributors to biological activity, so better rankings are illustrated by lower numbers (with a ranking of 1 being the best). Comprehensive results of these models can be found in Appendix B,

Table S2 and a workflow can be found in Scheme 2. In four datasets of the 18 generated, no cross-validated models could be produced. Three of these belonged to datasets obtained at low concentrations (0.01 mg/mL) injected to the mass spectrometer.



Scheme 2. Workflow for Untargeted Metabolomics Study in which Inactive *A. keiskei* Root Extract was Spiked with Known Antimicrobial Compounds. Biochemometric modeling results, and the impact of the of number of pools for chromatographic separation, concentration used for biological activity evaluation, and concentration injected into the LC-MS were evaluated. Additionally, the utility of data processing approaches, including data filtering and model simplification, were evaluated.

In datasets produced using chromatographic fractions separated into five (pools 2-1 through 2-5) or ten pools (pools 3-1 through 3-10), berberine and magnolol were the only constituents concentrated enough to contribute to biological activity. In the three-pool datasets (modeled using pools 1-1 through 1-3), cryptotanshinone was concentrated sufficiently to contribute to biological activity when pools were tested at a concentration of 100 µg/mL. As such, all models produced were expected to identify both berberine and magnolol as bioactive, but only the 3-pool datasets were expected to identify cryptotanshinone. Berberine was correctly identified among the top contributors to bioactivity (highest selectivity ratio) in 13 out of 14 models produced, 8 of which

identified berberine as *the* top contributor to biological activity. Magnolol was correctly identified as contributing the biological activity in all 14 models produced. Magnolol was identified among the top ten contributors to biological activity in only two out of fourteen models and was identified among the top twenty contributors in in ten of the remaining models. Cryptotanshinone, due to its low abundance, was only concentrated sufficiently to contribute to biological activity in the 3-pool set tested at 100 µg/mL. It was identified as the 19th top contributor to biological activity of this mixture when injected into the LC-MS at 0.1 mg/mL, but was not identified in the dataset assessed at 0.01 mg/mL.

Many problems in statistical analysis of metabolomics datasets arise because the number of samples (in this case, chromatographically separated mixtures) is typically greatly outnumbered by the variables analyzed (i.e. mass/retention time pairs) (256, 276, 287, 288). For example, our models compared between 9-30 samples (9, 15, or 30 samples for models produced using 3, 5, or 10 chromatographic pools) using 370 or 870 variables (mass/retention time pairs of models assessed via LC-MS at 0.01 mg/mL and 0.1 mg/mL, respectively). This low sample-to-variable ratio can lead to erroneous biological conclusions caused by correlation of nonactive to active metabolites under analysis (256, 276, 287, 288). In all models produced, numerous compounds were predicted to be active that were in fact components of the inactive botanical extract. It is important to note that without isolating each of these compounds and testing them individually, it is impossible to confirm their lack of bioactivity. However, to conservatively estimate the success of selectivity ratio models, they have been identified here as false positives. These false positives were of two types: those that co-varied with

spiked active compounds and those that did not. Co-varying false positives can be defined as compounds that were identified in the same pools, and with the same relative shifts in concentration, as active compounds. Non-co-varying false positives were identified as putatively active despite the fact they showed only minor variation across pools and did not share concentration shifts with active compounds. The identification of non-co-varying false positives is due to correlated noise, i.e. minor random variation in the bioassay data correlating to patterns in the concentration data (289). This is an important distinction because we aim to utilize this bioinformatics approach to guide the isolation process. While co-varying false positives will lead to the chromatographic separation of pools that possess active compounds (albeit not the compounds predicted), non-co-varying false positives may lead to the separation of a sample that will not yield an active compound.

To visualize the distinction between co-varying and non-co-varying false positives, five compounds found within the five-set are compared in relation to biological activity (Figure 28). Relative peak areas (based on percentage of the abundance across all pools) are displayed for each compound. In this example, berberine and magnolol (orange and blue bars, respectively), which were intentionally spiked in to the mixture, are responsible for the biological activity witnessed in pools 2-1 and 2-3, respectively. Additional ions are detected in the mixture (components of the original inactive botanical mixture) that co-vary with these active compounds in a way that makes their contribution to activity indistinguishable from true active compounds (represented by yellow and gray bars in Figure 28). For a mixture of truly unknown composition, these ions would qualify

as “false positives” and the analyst would not know if they or the actual known constituents were responsible for activity. A non-co-varying compound (light blue bar), is found in all pools under analysis at approximately equal concentrations, yet is still identified as a potential contributor to biological activity.

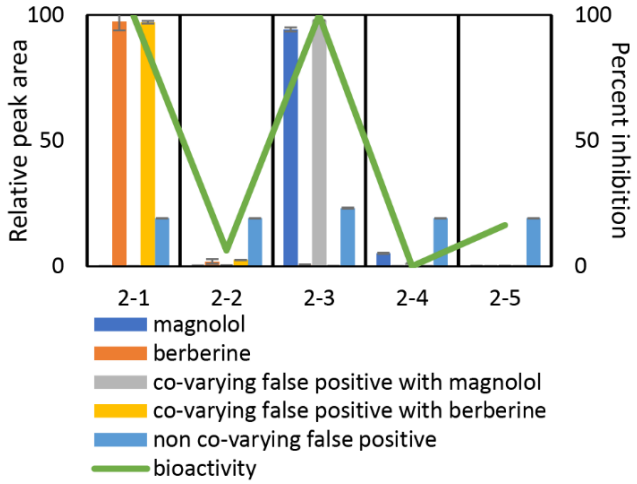


Figure 28. Relative Peak Area (Expressed as a Percentage of the Total Peak Area Detected Across Pools) of Berberine (Compound 1), Magnolol (Compound 2), and Selected “False Positives” Identified using Biochemometric Modeling Compared to Biological Activity Witnessed in Pools 2-1 through 2-5. Berberine and magnolol are responsible for the activity witnessed in pools 2-1 and 2-3, respectively. Co-varying false positives (yellow and gray bars) did not contribute to biological activity, but share the same abundance profiles as true active constituents across pools, and thus statistical models could not disentangle their contributions from those of the true bioactive constituents (berberine and magnolol). A non-co-varying false positive (light blue bar) is also illustrated. This component does not share abundance profiles with active constituents and is found at approximately equal abundance (±5%) across all pools. It represents correlated noise between biological activity and concentration data identified by the PLS model.

In the models produced, 2-18% of variables had selectivity ratios higher than 0, suggesting that variables in these subsets are likely to possess biological activity. Most of the false positives within these subsets were found in the same pools, and with the same relative shifts in concentration, as berberine and magnolol (representing between 43-85% of variables with selectivity ratios higher than 0 across all models produced, Appendix B,

Table S3). False positives that did not co-vary with berberine or magnolol were rarely a problem in datasets assessed at 0.1 mg/mL in the mass spectrometer. All models produced for low concentration datasets (0.01 mg/mL) had false positives that did not co-vary with known active compounds representing between 13-43% of variables with selectivity ratios greater than 0 (Appendix B, Table S3**)**. These findings illustrate that the low concentration datasets are more prone to overfitting and may lead to false biological interpretations.

Data acquisition and data processing parameters used to evaluate success of selectivity ratio models

Various types of data are collected to conduct complex metabolomics studies, particularly those involving biological activity, and each stage of data collection involves choices that may affect subsequent statistical analyses. Biological activity can be measured at a range of concentrations, and LC-MS data can be acquired using samples analyzed at different concentrations. High concentrations will allow more compounds to be detected by the mass spectrometer but may risk saturating the response of highly abundant or ionizable compounds. Low extract concentrations will be less likely to be subject to saturation, but low-abundance compounds contributing to activity may be overlooked if they are below the limit of detection for the LC-MS system. Finally, the number and chemical simplicity of chromatographic pools could also influence the metabolomics models.

We sought to evaluate the impact of the number of pools, bioassay concentration, and concentration analyzed by the mass spectrometer on the final biochemometric results.

To do this, we constructed models using different parameters and compared the resulting selectivity ratio rankings of berberine and magnolol. Berberine and magnolol were chosen because they were the only two added compounds that were concentrated enough following chromatographic separation to contribute to biological activity in all models tested. We also assessed the impact of number of pools, bioassay concentration, and concentration analyzed by mass spectrometry on the number of false positives, including false positives that co-varied with berberine, those that co-varied with magnolol, and those that did not co-vary with either active compound.

Effect of data acquisition parameters on selectivity ratio analysis

The models produced were built using ranked data, and as such, they do not meet assumptions of normality (290). Additionally, four of the eighteen subsets did not produce models, leading to a breaking of orthogonality. As such, we chose to use a partial least squares (PLS) analysis to assess the impact of the number of pools, bioassay concentration, and concentration injected into the LC-MS system on each of the result metrics (ranking of berberine, ranking of magnolol, false positives co-varying with berberine, false positives co-varying with magnolol, and non-co-varying false positives). The model generated to assess the variability among the selectivity ratio rankings of berberine explained 32.4% of the variability ($R^2 = 0.324$), suggesting that the number of pools included in the model, the bioassay concentration, and the mass spectral concentration have only a minor effect on the ability of selectivity ratio models to identify berberine as active. Similar results were found with selectivity ratio rankings of magnolol. Data acquisition parameters had a greater effect on the selectivity ratio

rankings of magnolol than berberine (R^2 = 0.484). The number of pools and the concentration tested in the bioassay did not have much impact on either model produced, and most of the variability was explained by concentration injected into the LC-MS, with high concentration datasets leading to better selectivity rankings. False positives co-varying with berberine were modeled using a 1-component model (R^2 = 0.627), and the number of false positives increased with increased concentration injected into the LC-MS. Interestingly, the false positives co-varying with magnolol were found to increase with the number of pools (R^2 = 0.901). Non-co-varying false positives increased with the number of pools and decreased with increasing concentration injected into the LC-MS and used in the bioassay (R^2 = 0.556).

Models produced using high concentrations in the LC-MS (0.1 mg/mL) were comprised of 870 unique ions. Of these 870 ions, a subset of ions, representing 2-5% of the total number of ions, had selectivity ratio rankings greater than 0. The low-concentration dataset (0.01 mg/mL), was comprised of 370 ions, and a subset containing 9-18% of the total number of ions possessed selectivity ratio rankings greater than 0. In all cross-validated models, between 14-34% of variables with selectivity ratios greater than 0 represented berberine or magnolol, including adducts and isotopes (Appendix B, Table S3). Our analyses revealed that datasets analyzed at higher concentrations analyzed in the LC-MS (0.1 mg/mL rather than 0.01 mg/mL) had improved selectivity ratio rankings for both berberine and magnolol, and also reduced the number of false positives that did not co-vary with active compounds (Appendix B, Table S2). These results suggest that saturation of highly abundant compounds (such as berberine) did not result

in a breakdown of linearity and allowed for the identification of active compounds. Models were made worse when assessed at lower concentrations, particularly for magnolol selectivity ratio rankings (Appendix B, Table S2**)**. We infer that at low concentrations, magnolol may be present at levels near or below the limit of quantification, skewing the linearity of the response and decreasing its contribution to the model. Low-concentration datasets appeared to be more prone to identifying correlated noise, as illustrated by the increased number of non-co-varying false positives (Appendix B, Table S2). Although there were more false positives that co-varied with berberine in the high concentration datasets, these numbers were small (1 or 2 false positives), and as such, the benefits of high concentration analysis outweigh the risk of false positives. Not only are high-concentration datasets less likely to identify non-co-varying false positives as active, they also provide a smaller pool of putative active compounds than those of low concentration datasets (2-5% versus 9-18%, Appendix B, Table S3).

Effect of data processing approaches on selectivity ratio analysis

Because of the immense complexity of botanical extracts, it is quite challenging to determine the number of metabolites present in a given sample (255, 276). Often, metabolomics datasets contain thousands of individually detected variables, whose signal intensities vary over a very large range, and may result from the detection of experimental artefacts (246, 258, 276). Data pre-treatment, filtering of chemical interferents, and model simplification tools may be critically important to enable extraction of relevant information from such datasets (276, 291). To explore this possibility in the context of natural products drug discovery, the impact of data

transformation, data filtering, and model simplification, as well as their second-order interactions, were assessed using data from the 10-pool set analyzed at 100 µg/mL in both the bioassay and by the LC-MS.

To measure the effects of data processing, we evaluated the selectivity ratio rankings of berberine and magnolol, as well as the occurrence of false positives, including those co-varying with berberine and magnolol and those that did not (Appendix B, Table S4). The six terms included in these models (data transformation, data filtering, model simplification, and second-order interactions) had excellent explanatory power in all models produced, explaining 95.2% of the variance of berberine selectivity ratio rankings, 99.6% of magnolol selectivity ratio rankings, 92.4% of false positives co-varying with berberine, 99.8% of false positives associated with magnolol, and 99.7% of the non-co-varying false positives. Depending on the combination of data processing approaches utilized, we found drastic changes in the selectivity ratio ranking of berberine (ranging from first to 23rd) and magnolol (ranging from 8th to 213th). A wide range was also witnessed for all categories of false positives (Appendix B, Table S4, Figure 29). These results suggest that data processing approaches are particularly important for extracting reliable information from metabolomics datasets.

Data transformation. It is common practice in metabolomics studies, particularly those utilizing mass spectrometric data, to subject data to a transformation procedure (241, 291). Because mass spectrometers are so sensitive in their ability to detect compounds at a wide range of concentrations, they are subject to errors caused by heteroscedastic noise in count data, in which error is proportional to the peak area (241,

291). As such, data transformation processes aimed to reduce the error associated with large peak areas are commonly employed (241, 291). Many metabolomics projects utilize, for example, a fourth-root transformation of variable peak areas to minimize the impact of heteroscedastic noise and reduce bias against highly abundant or ionizable compounds (132, 241, 292). Despite the popularity of this approach, our statistical analysis revealed that this transformation negatively impacted the ability of models to accurately predict active compounds. Models built using transformed data (Figures 29E-29H) gave berberine and magnolol worse selectivity ratio rankings than datasets using non-transformed data (Figures 29A-29D). There were also more false positives that did not co-vary with active compounds and that co-vary with berberine. Somewhat surprisingly, no false positives that co-varied with magnolol were detected in models that did not use transformed data. Likely, models that used transformed data were unable to identify magnolol as important for bioactivity, and as such, the compounds that co-varied with magnolol were not identified either. Because the non-transformed datasets were able to identify magnolol as active, the false positives associated with magnolol also increased. These results are counter to the findings of other studies (292). For example, while Arneberg et al. (292) found that the n^{th} root transformation positively impacted their models, our models using this transformation were unable to identify active constituents. These differences may be due to the differences in applications between these two projects. While Arneberg et al. (292) were assessing proteomics datasets, our datasets were focused on metabolomics-driven natural products discovery. In natural products discovery projects, low-abundant constituents that contribute to bioactivity may

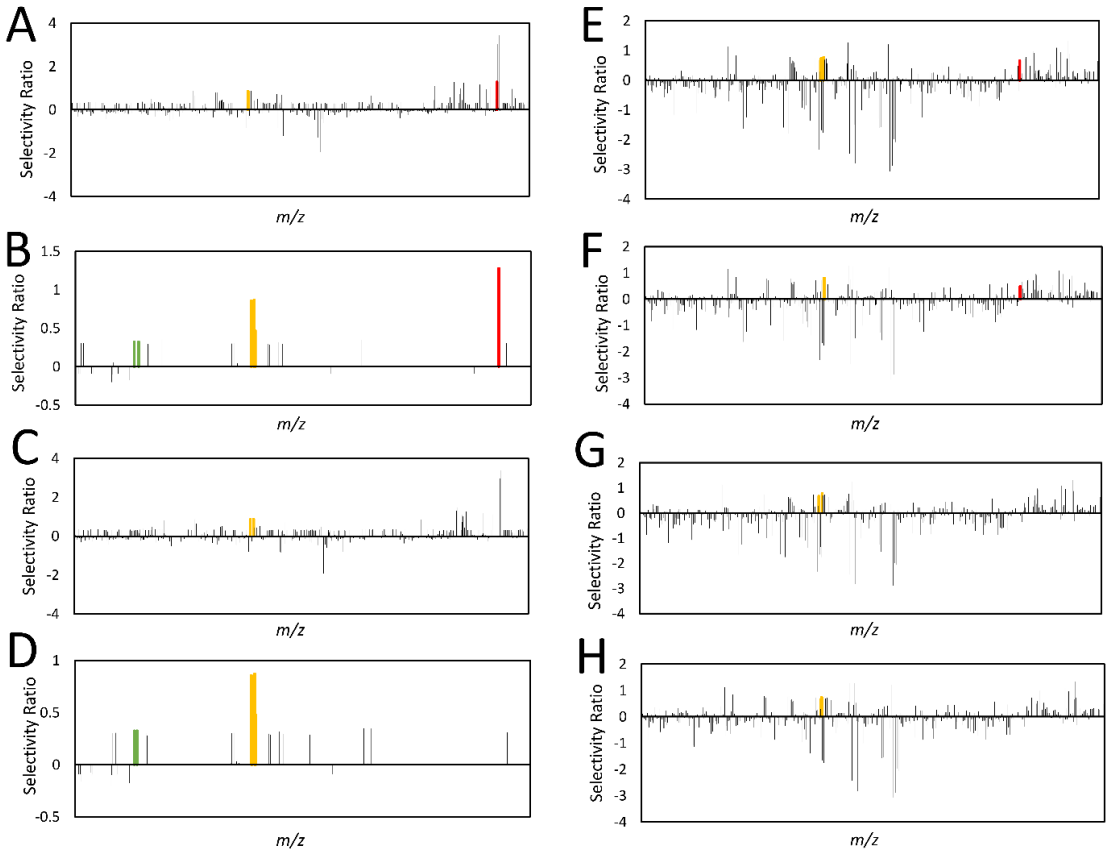


Figure 29. Comparison of Selective Ratios Produced with Different Data Processing Approaches. All models were derived from the 10-pool set analyzed at 0.1 mg/mL in the mass spectrometer using bioassay data at 25 µg/mL. *m/z*-retention time pairs (x-axis, high to low *m/z*) are plotted relative to their selectivity ratios (y-axis). The most positive selectivity ratios represent compounds with the highest explained to residual variance, and are predicted to be associated with biological activity. A series of identified features were associated with berberine and marked in yellow, including an $[M]^+$ ion at *m/z* 336.123 and retention time (RT) 2.96 min, an $[M]^+$ ion with *m/z* of 338.127 and RT of 2.961 min (containing two ^{13}C isotopes), an $[M]^+$ ion at *m/z* 339.129 min and RT 2.94 (containing three ^{13}C isotopes), and an $[M]^+$ ion at *m/z* 336.126 at RT 6.355 min. Two features were identified as associated with magnolol, and are marked in green, representing the $[M\text{-}H]^-$ ion at *m/z* 265.123 and ^{13}C isotope at *m/z* 266.127 at RT 5.756 min. Polysiloxane contaminants are marked in red. **29A.** No data processing approaches were used. **29B.** Model simplified using a percent variance cutoff, in which ions showing less than 1% peak area variance across samples (when compared to the most variable peak) were assigned a ratio of 0. **29C**. Model filtered using hierarchical cluster analysis (HCA), detailed in Caesar et al. 2018 (276) **29D.** Model simplified using percent variance cutoff and filtered with HCA. **29E.** Model produced using peak area data transformed with a fourth-root. **29F.** Model using transformed data and variance cutoff. **29G.** Model using transformed data and HCA filtering. **29H.** Model built with transformed data, filtered with HCA, and simplified using a percent variance cutoff. The model in Figure 4D has the fewest false positives and the best selectivity ratios for both berberine and magnolol, illustrating that its combination of data processing techniques are most suitable for this application.

be present in the upper parts per million or parts per thousand range (293), while protein biomarkers are often found in the lower parts per billion range (294, 295). A transformation to reduce the impact of major peaks compared to minor peaks may be helpful when the compounds of interest are likely to be extremely low in abundance, but not necessarily in the case of natural products discovery. Another potential reason for the negative impact of transformation on selectivity ratio models is that the fourth-root transformation is a nonlinear transformation, which may cause a breakdown in the linear relationship between active compound concentration and bioactivity.

Model simplification. The goal of this project is to identify active constituents from complex botanical mixtures, therefore, supervised methods using biological activity as the dependent variable should be used. Because the biological activity varies from sample to sample (Figure 26), the variables responsible for biological activity should also vary in concentration from sample to sample. To reduce the influence of variables that do *not* vary in concentration across pools on model interpretation, peak area variance was assessed. Variables were ranked according to their overall peak area variance between pools, and the variable with the highest variance was used as a reference. If variables contained an overall peak area variance that was less than 1% than that of the reference variable, it was assigned a selectivity ratio of 0. Datasets that were evaluated using this approach (Figures 29B, 29D, 29F, and 29H) had better selectivity ratio rankings for berberine and magnolol than those that did not (Figures 29A, 29C, 29E, and 29G). Additionally, there were fewer false positives that co-varied with berberine and that did not co-vary with active compounds in simplified models when compared to their non-

simplified counterparts. There were more false positives associated with magnolol in models that were produced using this simplification process, possibly because simplified models were better able to identify magnolol, and variables correlated with it, as important for biological activity.

Interaction between data transformation and model simplification. Multiple studies have been conducted to evaluate the influence of data processing treatments on subsequent data analysis, and have revealed that there are often complex interactions between the parameters used (292, 296). To optimize data treatment parameters, it is important to inspect interactions between processing steps. Indeed, our analyses also revealed a strong interaction between two data processing steps: data transformation and model simplification using a percent variance cutoff (Figures 29B and 29D). Models that did not use transformed data were better than their transformed counterparts at identifying berberine and magnolol as active only when model simplification using a percent variance cutoff was utilized. Transformed datasets were barely improved using this simplification method, likely because the data transformation minimized peak area variance between different ions. Models evaluated without data transformation and with a percent variance selectivity ratio filter (Figures 29B and 29D) showed enhanced selectivity ratio rankings for both berberine and magnolol. The selectivity ratio ranking for berberine in these models was 1^{st} or 2^{nd}, while all other models had selectivity ratio rankings between 17 and 23. The ranking of magnolol was 8^{th} or 9^{th} in models that were not transformed but were simplified using a percent variance selectivity ratio filter, while all other models had magnolol selectivity ratio rankings between 110 and 213. The

number of false positives that did not co-vary, as well as false positives co-varying with berberine, were also reduced. Again, the number of false positives co-varying with magnolol was increased in these datasets (Appendix B, Table S4**)**.

Data filtering using relative variance and hierarchical cluster analysis of triplicate injections. Often in mass spectrometry-based metabolomics, background noise and chemical contaminants are assumed to be consistent across samples. However, as illustrated in a recent study by the authors (276), this is not always the case. Chemical interferents originating from the analytical instrumentation itself (260, 261), including silica capillary contaminants and HPLC column packing materials, may be introduced differentially from injection to injection, in which case they will not be consistent across samples. Data filtering for removal of these contaminants from metabolomics datasets can improve quality and interpretability. This data filtering approach, when applied to the data collected herein, did not result in statistically significant changes to selectivity ratio rankings of berberine and magnolol, nor in the number of false positives identified (Appendix B, Table S4). However, in all models that did not go through this data filtering process, between one and four contaminants were incorporated into the model predictions. In one example, a known polysiloxane contaminant (271) was falsely identified as the top contributor to biological activity (Appendix B, Table S4, Figure 29B). Because many metabolomics studies rely on the assumption that compounds that vary in abundance from sample to sample may have biological importance, these types of contaminants are particularly important to identify and remove from metabolomics datasets.

Assessment of combination effects in unfractionated, spiked *A. keiskei* mixture

Many studies have shown that the observed biological activity of botanical mixtures may be due to the combined action of multiple constituents, which can interact additively, synergistically, or antagonistically (9, 10, 14, 18, 127). For the study conducted here, we hypothesized that such combination effects could be responsible for the large discrepancy in the predicted and observed activities for the spiked *A. keiskei* botanical extract (Figure 27). Specifically, we proposed that constituents of the 'inactive' botanical extract might mask or antagonize the antimicrobial activity of the antimicrobial compounds that had been spiked into it. To test this hypothesis, a checkerboard assay typically employed to assess synergy and antagonism in antimicrobial activity (18, 127, 297) was conducted in which purified berberine and magnolol were individually tested for antimicrobial activity in combination with a range of concentrations of the spiked *A. keiskei* mixture. The results of the synergy assay were illuminating, as illustrated in Table 9 and Figure 30. The spiked extract, when tested in combination with berberine, caused the minimum inhibitory concentration (MIC) of berberine to change from 75 µg/mL to 150 µg/mL and the IC_{50} to change from 29.5 µg/mL to 85 µg/mL (Figure 30A). Although these numbers may be suggestive of an antagonistic effect, using conservative ΣFIC indices, this effect was considered "noninteractive" (9). The spiked *A. keiskei* mixture had an even more notable impact on antimicrobial activity of magnolol (Figure 30B). The MIC of magnolol in combination with the spiked *A. keiskei* extract was increased to 25 µg/mL, when in isolation the MIC of magnolol was *four times lower* at 6.25 µg/mL. The

Table 9. Minimum Inhibitory Concentrations and Half Maximal Inhibitory Concentrations for Berberine (Compound 1) and Magnolol (Compound 2) Alone and in Combination with Spiked *A. keiskei* Extract. The MICs of berberine and magnolol in are consistent with previous reports (132, 284).

Treatment	MIC (µg/mL)	IC_{50} (µg/mL)	FIC index [a]
Berberine (**1**)	75	29.5	--
Berberine (**1**) + spiked *A. keiskei* extract [b]	150	85	3
Magnolol (**2**)	6.25	4.1	--
Magnolol (**2**) + spiked *A. keiskei* extract [b]	25	8.9	5
Spiked *A. keiskei* extract	>100 µg/mL	>100 µg/mL	--

[a] ΣFICs were calculated using the following equation: $\Sigma FIC = FIC_A + FIC_B = ([A]/ MIC_A) + ([B]/MIC_B)$, where A and B are the compounds/extracts tested in combination, MIC_A is the minimum inhibitory concentration of A alone, MIC_B is the minimum inhibitory concentration of B alone, [A] is the MIC of A in the presence of B, and [B] is the MIC of B in the presence of A.
[b] values expressed for magnolol and berberine's MIC/IC_{50} in combination with 100 µg/mL spiked extract.

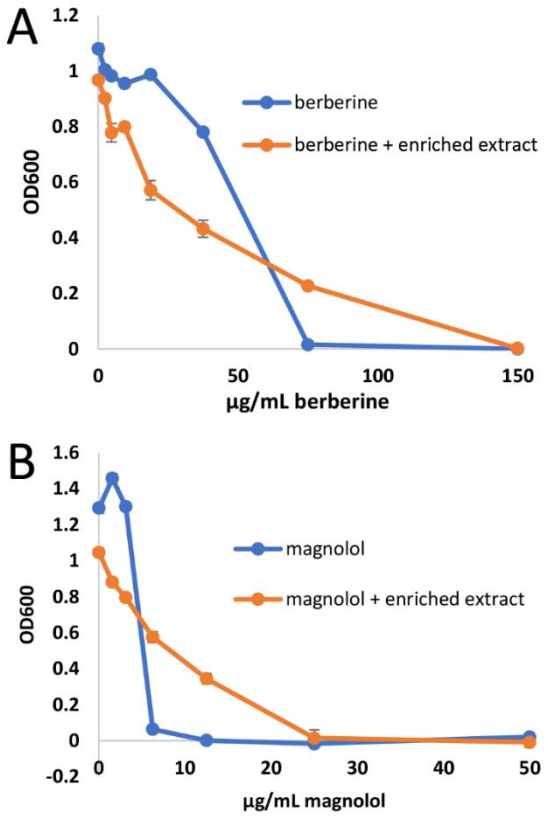


Figure 30. Comparison of Dose-Response Curves for Berberine (Compound 1) Alone and in Combination with 100 µg/mL Spiked Extract (A) and for Magnolol (Compound 2) Alone and in Combination with 100 µg/mL Spiked Extract (B). As indicated by the data shown here and the ΣFIC values in Table 1, the spiked extract antagonized the antimicrobial activity of the pure compounds. MIC values of compounds alone are consistent with previous reports (132, 284).

IC_{50} of magnolol was also impacted, and increased from 4.1 µg/mL in isolation to 8.9 µg/mL in combination with 100 µg/mL of the spiked mixture. The ΣFIC index for the magnolol/extract interaction was calculated to be 5, strongly indicating the presence of antagonists in the mixture. These results explain the mismatch in activity between our predicted and observed activity (Figure 27) and confirm the prediction that the mixture contains antagonists. Unfortunately, due to material limitations, identification and isolation of antagonists in the mixture was not pursued.

Assessing stage of fractionation and impact on assignment of bioactive constituents

Multiple rounds of fractionation improve selectivity ratio ranking of magnolol

Our analyses revealed that many compounds that co-varied with magnolol were incorrectly assigned as being bioactive. We anticipated that another round of fractionation and biochemometrics modeling would improve the selectivity ratio ranking of magnolol and eliminate some of these false positives. To this end, we separated three pools rich in magnolol (1-2, 2-3, and 3-5) with a second stage of chromatographic separation and evaluated their antimicrobial activity (Figure 31). The chromatographic separation of pool 1-2 yielded 11 sub-pools, pool 2-3 yielded 10 new sub-pools, and pool 3-5 yielded 7 new sub-pools. At 50 µg/mL, four of the new sub-pools caused complete inhibition of *S. aureus* (SA1199) (238) growth (Figure 31), while at 25 µg/mL, the most active sub-pool exhibited 60% inhibition.

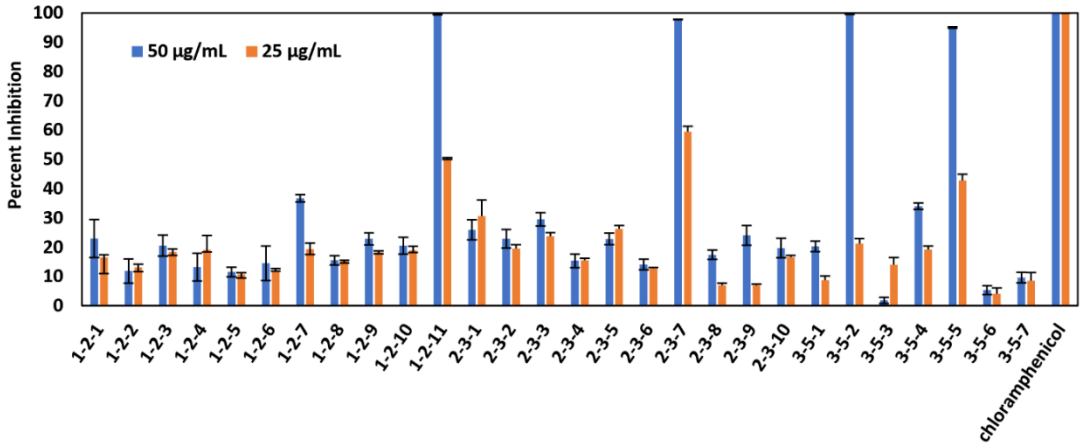


Figure 31. Biological Activity Data of Sub-Pools Resulting from Chromatographic Separation of Pools 1-2, 2-3, and 3-5, which Contained Active Concentrations of Magnolol. Growth inhibition of *Staphylococcus aureus* (SA1199) (238) relative to vehicle control was measured turbidimetrically using OD_{600} values. Data presented are the results of triplicate analyses ± SEM. The positive control chloramphenicol was tested at concentrations of 100 and 10 µg/mL.

Six new selectivity ratio models (two from each of the three new sets of sub-pools, assessed at 25 and 50 µg/mL) were produced using the sub-pool data from the second-stage fractionation (Appendix C, Figure S16), and these models were compared with the models generated from the previous round of fractionation (Appendix B, Table S5). The second-stage models had significantly higher selectivity ratio rankings for magnolol. Five of the six second stage models ranked magnolol between the 1st and 6th top contributors to biological activity (median ranking = 2), while their first stage counterparts ranked between 4th and 14th (median ranking = 13). Contrary to our predictions, the number of false positives were not affected by an additional round of fractionation.

Although the number of false positives found in the same chromatographic pools as magnolol were not affected, magnolol's contribution to the overall selectivity ratio models is more notable with second-stage pools. As an example, first- and second-stage

selectivity ratio models for the 10-pool set, analyzed at 0.1 mg/mL in the mass spectrometer, and assessed at 25 µg/mL are compared in Figure 32. Only the top 20 predicted contributors to biological activity are color coded. In this figure, red bars represent variables that co-varied with magnolol that were falsely identified among the top contributors to biological activity. Green bars represent magnolol and its associated masses (i.e. ^{13}C-isotopes). Blue bars are false positives that co-varied with berberine, and purple bars represent non-co-varying false positives. In Figure 32A, berberine and associated masses (yellow bars) are easily identifiable as putative active compounds, as are additional compounds that represent both co-varying and non-co-varying false positives. The green bars associated with magnolol are identified among the top twenty contributors to biological activity, but their relative magnitude is considerably smaller than many false positives. In Figure 32B the only false positives identified co-varied with magnolol, and magnolol's relative contribution to the model is improved. Berberine is not identified in this model because it was not present in pools selected for sub-fractionation.

Although false positives still prevail in the model predictions after additional rounds of fractionation, it is important to note that all the false positives in the top twenty contributors to activity in the second-stage model (Figure 32B) represent co-varying false positives. Because the impact of non-co-varying false positives is minimized by sub-fractionation, prioritization of pools for future chromatographic separation is more straightforward. Likely, an additional round of fractionation and modeling would improve this even further.

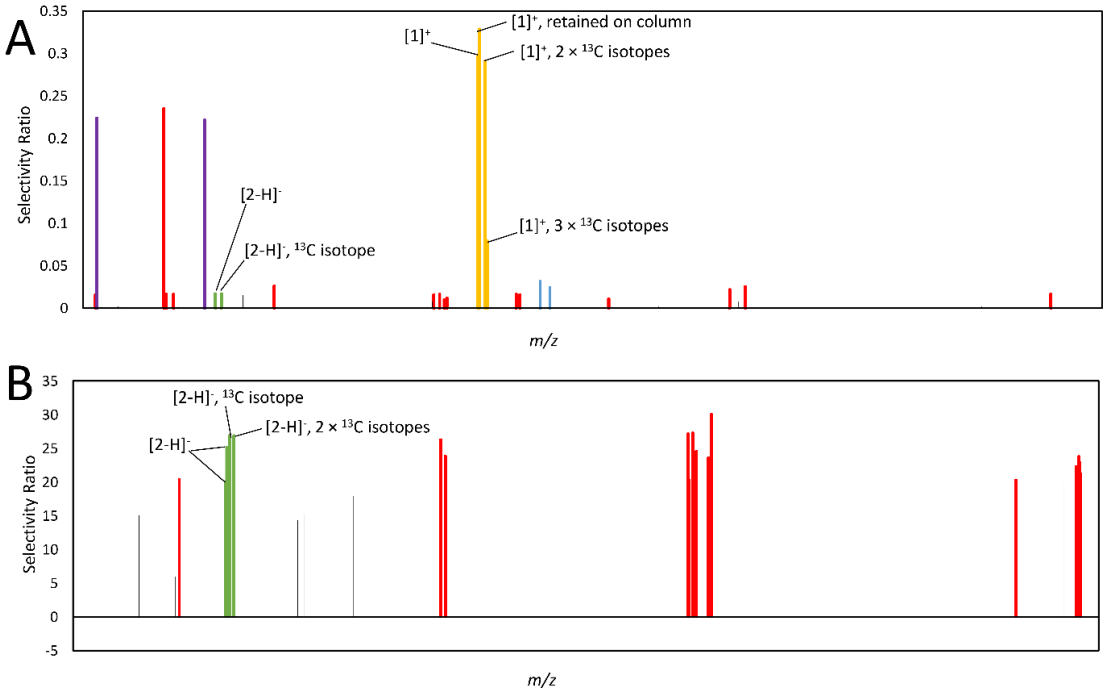


Figure 32. Models Produced using Pools 3-1 through 3-10 (32A) and 3-5-1 through 3-5-7 (32B) Analyzed at 0.1 mg/mL in the Mass Spectrometer and Assessed for Activity at 25 µg/mL. Features associated with berberine (compound **1)** are marked in yellow, and represent an $[M]^+$ ion at *m/z* 336.123 and retention time (RT) 2.96 min, an $[M]^+$ ion with an *m/z* of 338.127 and RT of 2.961 min (containing two ^{13}C isotopes), an $[M]^+$ ion at *m/z* 339.129 min and RT 2.94 (containing three ^{13}C isotopes), and an $[M]^+$ ion at *m/z* 336.126 at RT 6.355 min (RT difference due to column retention). Features associated with magnolol (compound **2**) are marked in green. In both **32A** and **32B** bars represent the $[M-H]^-$ ion at *m/z* 265.123 and ^{13}C isotope at *m/z* 266.127 at an RT of 5.756 min. Two additional associated ions, the $[M-H]^-$ ion at *m/z* 265.124 with an RT of 5.72, and the $[M-H]^-$ ion containing 2 ^{13}C isotopes at *m/z* 267.129 with an RT of 5.73 are found in **32B.** Co-varying false positives can be defined as compounds that were identified in the same pools, and with the same relative shifts in concentration, as active compounds. Non-co-varying false positives, on the other hand, were identified as putatively active but did not share concentration patterns with active compounds. In this figure, red bars correspond to variables co-varying with magnolol, blue bars represent false positives co-varying with berberine, and purple bars represent non-co-varying false positives.

These results are consistent with a recent study conducted in our laboratory exploring the use of biochemometrics and its ability to identify synergists in *Hydrastis canadensis* (18). With this project, three rounds of fractionation were required to produce a reliable selectivity ratio model. This model successfully identified known synergists in *H. canadensis* and revealed the activity of a previously undescribed compound (18). In

another study using biochemometrics and molecular networking to identify important constituents from *A. keiskei*, two rounds of fractionation data were required before antimicrobial compounds were identified (145). Thus, it appears that, as would be expected, biochemometric model predictions improve upon chromatographic separation. With the first set of models produced using complex first-stage pools, berberine was consistently identified among the top contributors to biological activity while magnolol was not. The pool containing the highest abundance of berberine from the first stage of chromatographic separation contained only 212 variables above the baseline, while the pool containing magnolol contains nearly twice as many compounds. However, after a second round of fractionation, the sub-pool containing the highest amount of magnolol only shows 310 ions above the baseline, making statistical modeling more efficient and less prone to data overfitting (Appendix C, Figure S17).

Multiple rounds of fractionation revealed an additional bioactive constituent previously masked by antagonists in the mixture

For the data shown in Figure 31, we can attribute the activity of sub-pools 1-2-11, 2-3-7, and 3-5-5 to magnolol, where magnolol was present at concentrations higher than its MIC (6.25 µg/mL) in sub-pools tested at 50 µg/mL (7.5 ± 1.2, 9.2 ± 0.4, and 10.5 ± 0.3 µg/mL for sub-pools 1-2-11, 2-3-7, and 3-5-5, respectively). However, sub-pool 3-5-2, which also inhibited growth of *S. aureus* (SA1199) (238) at 50 µg/mL, did not contain detectable levels of magnolol. Rather, this sub-pool was comprised almost entirely of another compound (93% purity based on LC-UV analysis, data not shown). We subjected this pool to an additional round of chromatographic separation, yielding randainal (**5**,

0.25 mg, 99% purity). Due to the structural similarity of randainal to magnolol, we propose that this compound did not originate from the *A. keiskei* root extract, but rather represented an oxidation product of magnolol. Indeed, raindainal was not detected in the unspiked *A. keiskei* extract used for these studies (data not shown).

Randainal was predicted by one second-stage model to be the fifth top contributor to biological activity. Nine false positives co-varied with randainal, and six false positives co-varied with magnolol. Three additional false positives were identified that did not co-vary with either of the active constituents. The discovery of randainal was illuminating, and highlights the importance of fractionation for identifying low-abundance antimicrobials *that may be masked by combination effects.* It appears that the presence of antagonists in the *A. keiskei* roots (Figure 30) masked the biological activity of randainal until it had been chromatographically separated from them. Although we were unable to test randainal in isolation for activity due to material limitations, its structural similarity to magnolol suggests that compounds present in the original *A. keiskei* mixture antagonized its activity in a similar way. Sub-pool 3-5-2 (93% randainal) was found to be active at 50 but not 25 μg/mL, which is likely the range of activity for randainal, although it is possible that minor constituents in the mixture also contributed.

The discovery of randainal also provided additional insight into models from the first round of data collection. Pool 3-6 possessed partial activity that was not explained by the four active compounds that we spiked into the mixture; however, this pool contained randainal, which likely contributed to the activity witnessed. Additionally, five of the original models identified randainal among the top contributors to biological

activity (Appendix B, Table 6). These masses were originally thought to be false positives that co-varied with magnolol.

Limitations and opportunities

Mass spectrometry is the analytical technology of choice in the metabolomics field because of its sensitivity to structurally diverse chemicals at a wide range of concentrations and ionization efficiencies. While mass spectrometry provides complex chemical profiles with the ability to reveal valuable scientific insights into various biological processes, it also is fraught with challenges. Especially when exploring complex biological organisms for unknown compounds, the analyst must contend with the fact that many variables detected may not represent compounds associated with the sample. Additionally, differences in ionization efficiencies of analytes detected can have major impacts on the statistical models produced. For example, we found that models produced when injecting higher concentrations into the mass spectrometer (0.1 mg/mL) were generally more informative. Although these models were at a higher risk for saturating the response of highly abundant compounds, they provided a more complete picture of true sample components. The low concentration datasets likely resulted in models that were skewed by highly abundant compounds, highly ionizable compounds, and noise. Low concentration models were less useful for identifying active compounds, and were also more prone to the inclusion of non-co-varying false positives due to correlated noise. Interestingly, data acquisition factors tended to impact the ranking of compounds identified as contributing to biological activity, but the identity of these candidates was relatively consistent.

Metabolomics datasets rely not only on the data acquired, but also upon the data pre-treatment and data processing steps utilized. Unlike data acquisition parameters, which affected the order but not identity of the top fifty ions produced, data processing parameters had a drastic impact on both order and identity of predicted bioactive constituents. Using a factorial design, we evaluated the effect of data filtering, data transformation, and model simplification steps on selectivity ratio analysis, and found that most of our models produced were unable to identify known active constituents and contained many putatively false relationships. One of the most substantial findings of this work was that data transformation, though commonly employed in metabolomics studies (241, 291), had a negative impact on subsequent statistical analyses. These results suggest that data processing protocols should be chosen carefully based on the goals of the project at hand and that commonly employed tools for one application may be unnecessary, or even detrimental, for other applications. We discovered that not only are individual pretreatment and processing steps influential (particularly model simplification using a percent variance cutoff and data transformation), but their interactions also have major impact on models produced. Finally, strategies to remove ions that do not represent real sample components are important for understanding the chemistry of the sample under analysis. Datasets that were not filtered using protocols described in a recent publication (276) contained false positive peaks associated with LC-MS equipment used for analysis. These peaks were often putatively identified as the top contributors to biological activity when the filtering approach was not utilized.

Even if all data acquisition and data processing parameters are optimized, there will likely be false negatives that are not incorporated into the model and false positives that are. For this experiment, we spiked four active compounds into a complex mixture. However, only two of these active compounds were concentrated enough to show biological activity. Alpha-mangostin, notably, was the most potent antimicrobial compound that we utilized; however, its low concentration in the pools that resulted from chromatographic separation prevented it from being detected as an active component of the original mixture. Cryptotanshinone was identified only in some of the models in which it was present at biologically relevant concentrations. Although this was not the case in this study, multiple rounds of fractionation may serve to concentrate low abundant active compounds enough to reveal their activity.

It is worth mentioning that the possibility of missing highly active compounds when they are present at low concentration is not only an inherent limitation of the biochemometric approach employed here, but of any bioassay guided fractionation experiment. It is almost always true that the analytical approach employed to profile natural product extracts and pools will be more sensitive than the biological assay employed to evaluate their activity. Thus, it is always possible for a detected compound to be falsely deemed "inactive" simply because it is present at levels too low to register a biological effect.

False positives are also a problem in biologically-driven metabolomics analysis. There will always be compounds that happen to be present in the same pools and at the same relative concentrations as true active constituents, so it is no surprise that inactive

compounds may be predicted to be active using a biochemometric approach. By utilizing optimized parameters for data processing and acquisition, it is possible to influence the *type* of false positives included in the model. False positives that are found in pools associated with biologically active constituents are less problematic than those that are not, because the fractionation process is guided by the predictions of the model. We have also found that antagonism can mask the activity of active compounds and distort metabolomics models. An additional round of fractionation allowed us not only to improve our identification of magnolol as active, but it also revealed an additional active compound, randainal, which was masked by combination effects. This compound was previously believed to be a false positive that was simply found in the same pools as magnolol. This finding suggests that many of the "false positives" we have counted in this study may not truly be false positives at all, but may represent active compounds whose activities have been distorted by combination effects.

Untargeted metabolomics is a tool for finding a needle in a haystack. For natural products drug discovery, the goal is often to identify bioactive "needles" in a haystack of thousands of metabolites. The studies described herein demonstrate that biochemometric approaches cannot necessarily identify the needle from the entire haystack, but rather, they can be applied to reduce the large haystack to a much smaller one that is likely to contain active compounds. Selectivity ratio analysis is an excellent tool to rank lead compounds in this smaller haystack and prioritize them for isolation. Effort is still required to purify the putative active compounds, assign their structures, and test them for biological activity. The studies presented herein demonstrate that such validation is very

necessary, given the likelihood of identifying false positives. However, the finite quantity of material available for subsequent isolation poses an inherent limitation that often stymies such validation.

Conclusions

The vast, largely unknown chemical landscape of botanicals is deeply rich, and although tools to understand the nature of their bioactive properties are improving, it is important to recognize that multivariate models are affected by a variety of biological, chemical, and analytical factors. Big data can be used to unveil valuable insights that are otherwise hidden to us. However, extracting information out of large datasets remains challenging. Despite this, we should not allow ourselves to be stagnated by imperfect or incomplete interpretations; rather, we should use our incomplete knowledge to generate hypotheses and strive to improve our interpretation and methods over time. This reality may remind us of statistician John Tukey's statement: "Far better an approximate answer to the *right* question . . . than an *exact* answer to the wrong question. Data analysis must progress by approximate answers, at best, since knowledge of what the problem really is will at best be approximate" (298). Although we may not find the exact answer to the question at hand, the effective management of large datasets gives us the ability to find better questions, recognize limitations, and follow up on predictions in an informed way.

Experimental Section

General experimental procedures

UPLC-MS analysis was conducted in both positive and negative modes using a Thermo-Fisher Q-Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific,

MA, USA) connected to an Acquity UPLC system (Waters Corporation, Milford, MA, USA). UPLC-MS analyses were completed using a reversed phase UPLC column (BEH C18, 1.7 µm, 2.1 × 50 mm, Waters Corporation, Milford, MA, USA). Each sample was analyzed in triplicate at concentrations of 0.1 mg/mL and 0.01 mg/mL in methanol (expressed as mass of sample per volume of solvent) with a 3 µL injection. Chromatographic separation was accomplished using a gradient comprised of water with 0.1% formic acid (solvent A) and acetonitrile with 0.1% formic acid (solvent B). The starting conditions were 90:10 (A:B) and held for 0.5 min. Over 0.5-8.0 min, the gradient was increased to 0:100 (A:B) and held at these conditions until 8.5 min. Over the next 0.5 min, starting conditions were re-established, and the gradient was held at 90:10 (A:B) from 9.0-10.0 min. Mass analysis (in both positive and negative modes) was completed over a *m/z* range of 150-1500. The settings were set as follows: capillary voltage -0.7 V, capillary temperature 310°C, S-lens RF level 80.00, spray voltage 3.7 kV, sheath gas flow 50.15, and auxiliary gas flow 15.16. A data-dependent method was used, and the four ions with the highest signal intensity were fragmented with HCD of 35.0.

Production of spiked botanical mixture with known antimicrobial compounds

The goal of this project was to evaluate the effectiveness of selectivity ratio analysis to identify known active (antimicrobial) compounds in an otherwise inactive mixture. Detailed information about the plant material, extraction, and simplification of this mixture can be found in Appendix A (Protocol S2). To prepare the spiked extract, a simplified and inactive *Angelica keiskei* Koidzumi extract (126.4 mg) was combined with four known antimicrobial compounds at different concentrations yielding 167.9 mg of the

spiked extract: berberine (**1**, 24.9 mg, 15% of extract mass), magnolol (**2**, 11.6 mg, 7% of extract mass), cryptotanshinone (**3**, 3.3 mg, 2% of extract mass), and alpha-mangostin (**4**, 1.7 mg, 1% of extract mass). This resulting mixture, containing both unknown compounds and known active compounds, was used as the test material for the experiments described herein.

Chromatographic separation experiments

The spiked *A. keiskei* root mixture was separated into three equal portions and reversed-phase HPLC was conducted. Each separation was conducted using the same gradient and column (Gemini NX reversed-phase preparative HPLC column, 5 µm C18, 240 × 21.20 mm; Phenomenex, Torrance, CA, USA) with a flow rate of 21.4 mL/min. The gradient began with 30:70 ACN:H_2O, after which it was increased to a ratio of 55:24 over 8 min. The gradient was then increased to 75:25 over two min and ramped up to 100% ACN for 28 min. The 100% organic gradient was then held for another two min to flush the column.

Each fractionation yielded 90 test tubes, which were divided evenly into sets containing three, five, or ten pools, facilitating assessment of the impact of chromatography and pool complexity on biochemometric analysis. The first set of pools consisted of three pools of 30 tubes each, the second set was made up of five pools with 18 tubes each, and the final set was ten pools of 9 tubes each. Each pool was dried under nitrogen before subsequent analysis. The three chromatographic separations of spiked *A. keiskei* root extract yielded eighteen pools, where 1-1 through 1-3 represent the samples

from the three-pool set, 2-1 through 2-5 represent the samples from the five-pool set, and 3-1 through 3-10 represent the samples from the ten-pool set.

Following the first round of biochemometric analysis, three pools were selected for a second round of chromatographic separation. The magnolol-rich pools, including the second pool from the three-pool set (pool 1-2), the third pool from the five-pool set (pool 2-3), and the fifth pool from the ten-pool set (pool 3-5). These pools were subjected to another round of reversed-phase HPLC. All pools were separated using a gradient comprised of acetonitrile and water through a Gemini NX reversed-phase preparative HPLC column (5 µm C18, 240 × 21.20 mm; Phenomenex, Torrance, CA, USA) with a flow rate of 21.4 mL/min. Pool 1-2 was separated using a gradient beginning with 45:55 ACN:H_2O and increasing to 60:40 ACN:H_2O over 30 min after which it was flushed with 100% acetonitrile for 10 min. Pool 2-3 was separated into ten sub-pools using a gradient increasing from 60:40 to 70:30 ACN:H_2O over 25 min and ending with a 10 min flush of 100% acetonitrile. Finally, pool 3-5 was separated into 7 sub-pools (3-5-1 through 3-5-7) with an isocratic gradient of 60:40 ACN:H_2O held for 30 min before a 10 min flush of 100% acetonitrile. Sub-pool 3-5-2, collected from 9-10 min, was subjected to a final round of reversed-phase HPLC through a Phenomenex Gemini-NX reversed-phase analytical column (5 µm; 250 × 4.6 mm) with a 35 min gradient of ACN:H_2O starting at 30:70 and increasing to 70:30 following which it was increased to 100:0 for 5 min. Randainal (**5**) was collected from 20-20.5 min (0.25 mg, 99% purity). NMR spectra were collected using an Agilent 700 MHz spectrometer (Agilent Technology) or a JEOL ECA-500 MHz spectrometer (JEOL, Peabody, MA, USA).

*Randainal (compound **5**)*: yellow, amorphous powder; HRESIMS *m/z* 279.1028 $[M-H]^-$ (calculated for $C_{18}H_{15}O_3^-$, 279.1021). Fragmentation patterns matched predicted patterns as well as previously reported fragments from the literature (299) (Appendix C, Figure S18); ^{1}H NMR (700 MHz, CD_3OD) *δ:* 3.34 (2H, d, J=6.3 Hz, H_2-7'), 4.57 (2H, s, OH), 5.00 (1H, ddd, J=10, 2, 1 Hz, H-9'$_a$), 5.06 (1H, ddd, J=17, 2, 1 Hz, H-9'$_b$), 5.98 (1H, ddt, J=17, 10, 6.7 Hz, H-8'), 6.58 (1H, dd, J= 15.6, 8 Hz, H-8), 6.81 (1H, d, J=8.2 Hz, H-5'), 6.86 (1H, d, J=8.4 Hz, H-5), 7.01 (1H, dd, J=8.2, 2.2 Hz, H-6'), 7.11 (1H, d, J=2 Hz, H-2'), 7.52 (1H, dd, J=8.5, 2.2 Hz, H-6), 7.55 (1H, d, J=2.3 Hz, H-2), 7.63 (1H, d, J=15.6 Hz, H-7), 9.52 (1H, d, J=7.9 Hz, H-9) (Appendix C, Figure S19). To assign shifts corresponding to protons in the aromatic rings, HSQC data (700 HMz, CD_3OD) were used to identify the correlation between H-2 and C-2 (Appendix C, Figure S20) and HMBC data (700 HMz, CD_3OD) were used to identity correlations between C-2 and H-7 and H-6 (Appendix C, Figure S21). Previous literature reports on this compound were completed in acetone-$d_{6\,(300)}$. To confirm the identity of this compound, we ran an additional ^{1}H NMR (500 MHz, acetone-d_6) whose chemical shifts matched literature values (Appendix C, Figure S22) (300).

Antimicrobial assay

To assess antimicrobial activity, a broth microdilution assay was completed for each pool using a laboratory strain of *Staphylococcus aureus* (SA1199) (238). Assays were conducted using Clinical laboratory Standards Institute (CLSI) standard protocols (236). Cultures were grown in Müeller-Hinton broth (MHB) from an isolated colony and diluted to 1.0×10^5 CFU/mL calculated using absorbance at 600 nm (OD_{600}) values.

Because one of our goals for this project was to assess the impact of bioassay data format on biochemometric results, a full dose response curve was completed for each pool and each known antimicrobial compound. Stock solutions were prepared in DMSO and diluted with MHB so that final concentrations in test wells would contain 2% DMSO. Using these stock solutions, samples were screened in triplicate at concentrations ranging from 0-100 µg/mL in MHB (or 0-150 µg/mL in the case of berberine). The 28 sub-pools produced during the second round of fractionation were screened for bioactivity testing at two concentrations: 50 and 25 µg/mL. Chloramphenicol was used as a positive control. Each well was inoculated with bacteria (at 1.0×10^5 CFU/mL) and incubated for 18 hours at 37 ºC. After incubation, OD_{600} was calculated using a Synergy H1 microplate reader (Biotek, Winooski, VT, USA) and used to calculate the growth inhibition of *S. aureus* by the pools and/or compounds tested. Minimal inhibitory concentrations (MICs) were calculated for each of the known compounds, defined as the concentration at which there was no statistically significant difference in OD_{600} values between the negative control (wells containing broth and samples but no bacteria) and the treated sample. Dose response curves were produced using a four-parameter logistic model in SigmaPlot (v.13, Systat Software, San Jose, CA, USA).

Synergy assessment

Antimicrobial checkerboard assays using a broth microdilution method (127, 297) were conducted to assess the effect of the spiked extract on the antimicrobial efficacy of berberine and magnolol. The *A. keiskei* extract, spiked with berberine, magnolol, cryptotanshinone, and alpha-mangostin, was tested in combination with berberine or

magnolol, with the spiked *A. keiskei* extract and magnolol ranging in concentration from 1.56-100 µg/mL, and berberine ranging from 2.34-150 µg/mL. The vehicle control was comprised of 2% DMSO in Müeller-Hinton broth. The fractional inhibitory concentration index (ΣFIC) for each combination of compounds was calculated using equation 1 (127):

$$\Sigma FIC = FIC_A + FIC_B,$$

$$\text{Where } FIC_A = [A]/MIC_A, \text{ and } FIC_B = [B]/MIC_B \qquad \text{(equation 1)}$$

A and B are the compounds/extracts tested in combination, MIC_A is the minimum inhibitory concentration of A alone, MIC_B is the minimum inhibitory concentration of B alone, [A] is the MIC of A in the presence of B, and [B] is the MIC of B in the presence of A. To minimize the risk of misinterpretation of data, which is common in interaction studies (9, 23, 39, 59, 301-303), we have chosen conservative values to assign combination effects as recommended in the review by van Vuuren and Viljoen (9). For the purposes of this project, synergistic effects are defined as interactions having an ΣFIC ≤ 0.5, additive effects have an ΣFIC between 0.5 and 1.0, non-interactive effects have ΣFIC values between 1.0 and 4.0, and antagonistic effects have ΣFIC values ≥ 4.0.

Quantitative analysis of known compounds and contribution to biological activity

Concentrations of known active compounds berberine, magnolol, cryptotanshinone, and alpha-mangostin were determined using LC-MS. An external calibration curve of each standard compound (with final concentrations ranging from 0-50 µg/mL in methanol) was produced to identify the linear range of the calibration curve. Each sample was re-suspended in methanol to a concentration of 0.1 mg/mL and

analyzed as described in *General Experimental Procedures*. Concentrations were calculated from the relevant calibration curve based on the peak area of the relevant selected-ion chromatogram for each compound in each sample. Antimicrobial dose-response curves of each compound tested in isolation were used to determine which pools possessed biologically relevant concentrations.

Statistical analysis

Baseline correction/MZmine parameters

LC-MS datasets acquired in both positive and negative modes were individually analyzed, aligned, and filtered using MZMine 2.21.2 software (http://mzmine.sourceforge.net/) (239). Raw data files (including triplicate analyses of each sample) were uploaded into MZMine for peak picking. Chromatograms were built for all *m/z* values having peaks lasting longer than 0.1 min. The spiked extract was subjected to two stages of fractionation (Appendix C, Figure S16). The first-stage models were produced using pools 1-1 through 3-10. Sub-pools used to produce second-stage models were generated by sub-fractionating pools 1-2, 2-3 and 3-5, and are labeled 1-2-1 through 3-5-7. Modeling completed with the first set of pools were produced using the following peak detection parameters: noise level (absolute value) of 2.0×10^6 (positive mode, 0.1 mg mL^{-1} samples), 1.0×10^7 (positive mode, 0.01 mg mL^{-1} samples), and 1.0×10^6 (negative mode, both 0.1 mg mL^{-1} and 0.01 mg mL^{-1} samples). Models produced using the second set of pools (1-2-1 through 3-5-7) resulting from chromatographically separating magnolol-rich pools (pools 1-2, 2-3, and 3-5) were assessed at 0.1 mg/mL. For these data, the noise level was set to 2.0×10^6 for both positive and negative modes. For

all modeling datasets, the *m/z* tolerance was set to 0.0001 Da or 5 ppm, and the intensity variation tolerance was set to 20%. Peaks were aligned if they were both within 5 ppm *m/z* from one another and eluted within a 0.2-min retention time window. Data consisting of *m/z*, retention time, and peak area, for both negative and positive ions was imported into Excel (Microsoft, Redmond, WA, USA) and combined as a single peak list. Biological data were added as percent inhibition of bacterial growth at 25, 50 and 100 µg/mL. Data matrices for each sample subset (containing different pool numbers, mass spectral concentrations, and biological activity data) were independently imported into Sirius version 10.0 (Pattern Recognition Systems AS, Bergen, Norway) (240) for statistical analysis.

Hierarchical cluster analysis and chromatograph visualization

Hierarchical clustering analysis was conducted on each data subset using Sirius version 10.0 (Pattern Recognition Systems AS, Bergen, Norway) (240, 269). Briefly, samples were analyzed using an average-linkage algorithm (270) to cluster objects based on chemical similarity. A dataset was considered clustered effectively only when triplicate injections of the same sample were linked to one another before being linked to other samples. If triplicates did not show this expected trend, spectral variables were inspected for each set of triplicates. Variables showing high peak area variability *within* triplicate injections, as well as their associated isotopes, in-source fragments, and clusters, were removed. Datasets were also produced that did not include this filtering process to assess the importance of this process on subsequent selectivity ratio analysis. For a more detailed description of this approach, see Caesar et al. 2018 (276).

Selectivity ratio analysis

Selectivity ratios were generated with Sirius version 10.0 statistical software (Pattern Recognition Systems AS, Bergen, Norway) (240, 269). As part of the goals of this project, we sought to assess the impact of various data transformation and filtering approaches on the resulting biochemometric analysis. Before analysis, peak area data were transformed using a fourth-root transformation to reduce heteroscedastic noise (241). Additional data subsets were produced in which the data were not transformed. Each subset was subjected to internally cross-validated PLS analysis using 100 iterations and a significance level of 0.05. Algorithms internal to the Sirius statistical software were computed, resulting in selectivity ratio plots that identified candidate compounds associated with biological activity. As a final filtering step, each variable within each dataset was assessed, and those showing lower than 1% peak area variance across samples were assigned a selectivity ratio of 0 in order to reduce the effect of correlated noise from the datasets. This resulted in more simplified selectivity ratio plots which were compared to plots that did not include this filtering step.

Statistical comparison of protocols

Partial least squares regression followed by target projection (269) and calculation of selectivity ratios (149) was used for calculating all models predicting biological activity from mass spectral profiles. Double cross validation (288) was used to determine the number of PLS components for each model.

For assessing the impact of data acquisition protocols (pool number, bioassay concentration, and mass spectral concentration) on the ranking of the bioactive

candidates, we did PLS regression with these variables and their interactions as explanatory variables in models predicting ranking of berberine and magnolol, and the number of false positives identified in the models. Similarly, the effects of 4^{th} root transformation, data filtering, 1% variance cutoff and their 2-factor interactions on the ability to reveal and rank bioactive compounds in the mass spectral data was assessed by calculating regression models with these variables and their interactions as explanatory variables.

Acknowledgements

This research was supported by the National Center for Complementary and Integrative Health of the National Institutes of Health under grant numbers 5 T32 AT008938 and 1R01 AT006860. Additionally, the authors would like to acknowledge Richo Cech for his provision of plant material, Dr. Alexander Horswill for the provision of microbial strains, and Sonja Knowles for her assistance with NMR analysis. Mass spectrometry analyses were conducted in the Triad Mass Spectrometry Facility at the University of North Carolina at Greensboro (https://chem.uncg.edu/triadmslab/). This work was performed in part at the Joint School of Nanoscience and Nanoengineering, a member of the Southeastern Nanotechnology Infrastructure Corridor (SENIC) and National Nanotechnology Coordinated Infrastructure (NNCI), which is supported by the National Science Foundation (Grant ECCS-1542174).

CHAPTER VI

SIMPLIFY: AN INTEGRATED METABOLOMICS APPROACH TO IDENTIFY ADDITIVES AND SYNERGISTS FROM COMPLEX MIXTURES

This chapter has been submitted to the journal Proceedings of the National Academy of Sciences and is presented in that style. Caesar, L.K., Nogo, S., Naphen, C.N., Cech, N.B.

Caesar, L.K. conceived of the idea for this project, and assisted with chromatographic separation, compound isolation, mass spectrometry analysis, and bioassay analysis. Caesar, L.K. produced all statistical models and interpreted data. Nogo, S. helped with sample preparation, chromatographic separation, mass spectral analysis, and bioassay data collection. Naphen, C.N. assisted with chromatographic separation and biological testing. Cech, N.B. assisted in the development of the research project and provided edits and suggestions throughout manuscript preparation.

Introduction

Analysis of complex mixtures is an important topic of scientific research, providing insight into many biological processes and interactions. A complex mixture under study, whether it be an environmental pollutant such as cigarette smoke, a microbial community collected from the deep ocean, or a botanical medicine, is frequently reduced to the contributions of its individual constituents (14, 18, 304, 305). Very often, however, the activity of one constituent may be affected by the presence of other compounds in the mixture (8, 14, 18, 127, 304, 305). While it is true that individual constituents of a complex mixture may contribute to their biological activity, our understanding of the mixture's activity as a whole often remains incomplete.

Natural products have evolved complex biosynthetic pathways to develop chemical defenses against pathogens, and as such, the diverse combinations of compounds produced could be harnessed as antibacterial therapeutics (221, 222, 224-227). Particularly because pathogenesis of antimicrobial-resistant infections is often achieved through multi-factorial mechanisms (306), phytotherapies that owe their activity to the combined action of multiple constituents may offer important treatment opportunities (8). These combination effects can result from mixtures possessing synergistic, additive, or antagonistic activity (9, 15, 16). The specific types of interactions possible within a complex botanical extract are numerous, and may involve the defense of an active substance from enzymatic degradation (61, 307, 308), inhibition of multi-drug resistance mechanisms (15, 127), modification of transport across cell membranes (61), and improvement of bioavailability (309). The presence of combination effects such as these may lead to the loss of biological activity when the mixture is reduced to its individual constituents in isolation (10, 14, 17, 310).

Most natural product discovery efforts are geared towards identifying single active constituents as leads for drug development (220). While this approach has been undeniably useful in identifying important pharmaceutical drugs such as taxol and camptothecin (311), combination effects are often overlooked with isolation-based approaches (14, 18). This is a problem because many natural products are used therapeutically as mixtures; thus, it is of interest to know how such mixtures act in their complex form. Furthermore, if the combination of compounds responsible for the activity of a mixture is known, it is possible to rationally design more effective mixtures

that may be therapeutically useful. Several methodologies have been reported for identifying combinations of compounds contributing to the bioactivity of natural product mixtures (131, 240, 247, 312). One such approach, termed synergy-directed fractionation, tests fractions of a mixture in combination with a known antimicrobial (14). Synergy-directed fractionation avoids the problem of overlooking compounds that potentiate activity but are inactive in isolation. However, like traditional bioactivity-guided approaches, synergy-directed fractionation is predisposed towards compounds that are most easily isolated (18). Consequently, the synergists identified may only represent a portion of the actual constituents involved in the biological effect under study.

Recently, effective bioinformatics tools have been developed that can integrate biological and chemical datasets to predict which constituents of a mixture possess biological activity (18, 132). Several reports have been published illustrating the applications of biologically-guided metabolomics studies (so-called "biochemometrics" approaches) to identify putative active compounds and synergists from natural product mixtures (18, 131, 132, 145). Until this point, however, these approaches cannot predict whether mixture components are interacting synergistically, additively, or antagonistically without purifying compounds and testing them in isolation. Additionally, existing assays capable of disentangling synergy from additivity are time consuming and require considerable quantities of material (14, 18). Thus, new tools are needed to enable the efficient identification of constituents that interact to achieve biological activity.

The goal of this work was to develop a predictive approach to prioritize the isolation of mixture constituents that interact synergistically, additively, or

antagonistically. As a case study, we utilized the botanical *Salvia miltiorrhiza* (Chinese red sage or danshen). This botanical has been employed for medicinal purposes for over 2000 years in China (313-318) and remains one of the most popular traditional medicines in use today (313, 316, 317). Over 70 unique constituents have been identified in *S. miltiorrhiza*, making it an excellent model system for developing improved analytical methodologies. One of the most abundant constituents of *S. miltiorrhiza* is cryptotanshinone, a compound that demonstrates antibacterial activity against a broad range of bacteria, both alone (313, 314, 317) and in combination with existing antibiotics (313, 314). While the activity of *S. miltiorrhiza* constituents have been tested in isolation, little is known about the activity of *S. miltiorrhiza* extracts as a whole (317). With this study, we demonstrate the effectiveness of Simplify, an approach combining biological activity studies with metabolomics models towards the identification of synergists and additives that interact to exert combined effects. A unique strength of the approach we develop is the ability to characterize the nature of the interactions that occur prior to isolation. The Simplify approach is relevant beyond the field of natural products and could prove useful to researchers investigating the biological activity of mixtures in fields ranging from toxicology to pharmacology to drug discovery.

Results

The Simplify approach reveals a mismatch between predicted and observed activity and enables prediction of mixture constituents responsible for combination effects

The general concept for the Simplify approach to identify synergists, antagonists, or additives is outlined in Figure 33. With this approach, the mixture is

chromatographically separated into a series of fractions, and the activity of these fractions is predicted based on the quantity of a known active constituent, which is either naturally present or spiked into the mixture. An activity index (equation 1) is then calculated for each fraction as follows:

$$\textit{Activity Index} = (\textit{Actual Activity / Predicted Activity}) \times 100 \qquad \text{(equation 1)}$$

which provides a quantitative metric to explain how much each fraction enhances or suppresses the activity of the known constituent. Because the activity index represents the percent enhancement when actual activity is ratioed to predicted activity, fractions with an activity index >100 have a potential to contain either synergists or additives, while fractions with an activity index of <100 may contain antagonists. However, it is not possible using the activity index alone to distinguish between synergy and additivity. To make this distinction, fractions with sufficient material showing larger than a 10% mismatch between predicted and actual activity are subjected to checkerboard assays (127), where specific combination effects can be identified based on the calculation of a fractional inhibitory concentration (ΣFIC)**.** The ΣFIC is calculated using equation 2 (127):

$$\Sigma FIC = FIC_A + FIC_B,$$

$$\text{Where } FIC_A = [A]/IC50_A, \textit{ and } FIC_B = [B]/IC50_B \qquad \text{(equation 2)}$$

In this equation, A and B are the samples tested in combination, $IC50_A$ represents the IC_{50} of A in isolation (calculated using a four-parameter logistic model), $IC50_B$ represents the IC_{50} of B in isolation, [A] is the IC_{50} of A in the presence of B, and [B] is

the IC_{50} of B in the presence of A (127). Several review papers have been published on the use and interpretation of isobolograms and ΣFICs for evaluating combination effects (8-10, 17, 61). In short, an isobologram is a plot where each x,y data pair corresponds to a combination of concentrations at which a desired activity is achieved (e.g. 50% inhibition of bacterial growth). The shape of the isobologram indicates whether the interaction is synergistic, additive, or antagonistic. The ΣFIC index (equation 2) is a quantitative measure that can be calculated using the same data (127). For the purposes of this study, an ΣFIC cutoff < 0.5 is indicative of synergy, an ΣFIC between 0.5 and 1.0 is indicative of additivity, an ΣFIC between 1.0 and 4.0 indicates indifference or no interaction, and an ΣFIC > 4.0 indicates antagonism.

Once ΣFICs have been calculated, the ΣFIC values are used to sort fractions by the type of combination effect they exhibit. As a final step, activity indices (equation 1) are used to guide biochemometric analyses and identify specific mixture constituents that contribute to the activity of the mixture. This application of the activity index is what makes the Simplify process unique as compared to other approaches. The biochemometric analysis of data enables identification (based on detected *m/z* and retention time) of specific mixture constituents that are expected to exhibit interactions. These compounds can then be prioritized for isolation and structure elucidation (using classical MS and NMR based approaches) and tested in combination to confirm the predicted activities.

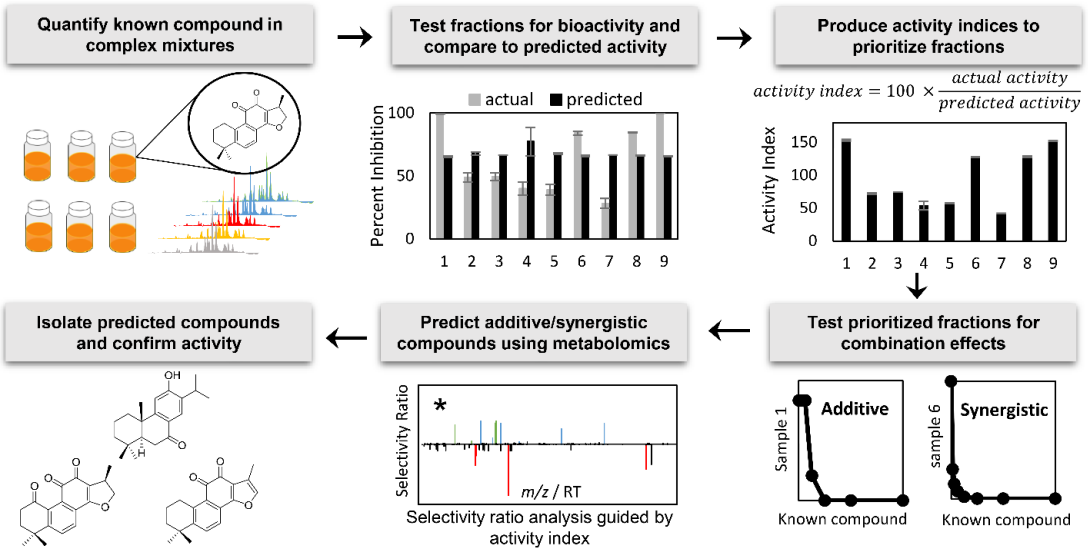


Figure 33. Workflow for the Simplify Approach. First, a known active compound is quantified in each complex mixture. The predicted activity based on the concentration of the known compound is then compared to the actual activity of each fraction, and an activity index is calculated to identify which fractions enhance/suppress the activity of the known active compound. Fractions showing a mismatch greater than 10% are prioritized for follow up testing, where additivity is disentangled from synergy. Activity indices (equation 1) are then used to produce selectivity ratio models predicting which constituents within the complex mixtures are responsible for additive or synergistic effects. Variables with high selectivity ratios are predicted to contribute to combination effects. Predicted active compounds are prioritized for isolation, following which predicted activities are confirmed.

The activity of cryptotanshinone is enhanced by both additives and synergists in complex *S. miltiorrhiza* mixtures

The first four steps of the Simplify approach (Figure 33) were employed at each stage of the fractionation process (Appendix C, Figure S23). With this approach, *S. miltiorrhiza* was extracted and fractionated with column chromatography, and a known antimicrobial compound cryptotanshinone (compound **1,** Figure 34) was quantified in each fraction. The predicted activity, based on the dose-response curve for pure cryptotanshinone, was compared with the observed biological activity to identify fractions that possessed less or more antimicrobial activity than could be explained based

on the measured concentration of cryptotanshinone. These fractions were prioritized for full checkerboard assays where specific combination effects could be identified.

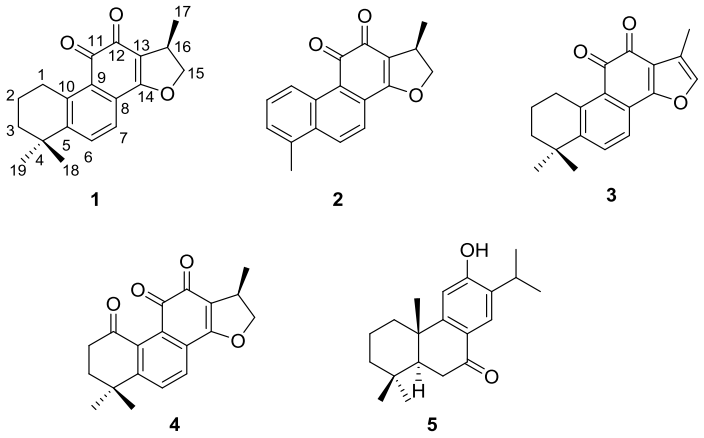


Figure 34. Compounds Identified from *S. miltiorrhiza* Utilized for this Study. Compounds 1-5 correspond to cryptotanshinone, dihydrotanshinone I, tanshinone IIA, 1-oxocryptotanshinone, and sugiol, respectively.

The first round of fractionation yielded eight fractions (SM-1 through SM-8), three of which possessed enhanced antimicrobial activity that was not explained by cryptotanshinone (fractions SM-1, SM-3, and SM-5, Figure 35A). To determine whether this enhancement in activity was synergistic or additive, checkerboard assays (127, 297) were conducted, in which a series of cryptotanshinone dose-response curves were collected in the presence of varying concentrations of the fraction under study. Using the data from these assays, isobolograms were plotted and fractional inhibitory concentration (ΣFIC) indices were calculated (Figure 35B-35D). Both SM-1 and SM-3 had ΣFIC values < 0.5, indicating that synergists were present in these mixtures (Figure 35B and 35C). Fraction SM-5 had an ΣFIC of 0.75, illustrating that additives rather than synergists contributed to the mismatch witnessed using this approach.

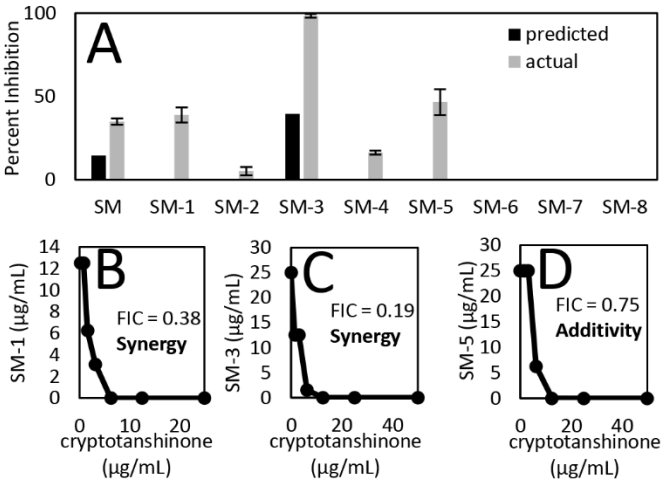


Figure 35. Comparison of Predicted to Actual Activity. (A), where black bars represent the antimicrobial activity of each fraction due to cryptotanshinone (predicted using peak are of cryptotanshinone and dose response curves of cryptotanshinone alone) and gray bars represent the actual activity of the fraction measured at 10 µg/mL (mass of extract per assay volume). Cryptotanshinone was used as a positive control, and its MIC (25 µg/mL) is consistent with previous reports (285). Fractions SM-1, SM-3, and SM-5 showed a mismatch between predicted and observed biological activity and were prioritized for synergy testing in combination with cryptotanshinone. The resulting isobologram of SM-1 shows synergy with an ΣFIC of 0.38 (**B**). SM-3 is synergistic with an ΣFIC of 0.19 (**C**), and the isobologram of SM-5 shows additivity with an ΣFIC of 0.75 (**D**).
ΣFICs were calculated using equation 2: $[A]/IC_{50}A + [B]/IC_{50}B = \Sigma FIC$, where $IC_{50}A$ is the IC_{50} of cryptotanshinone alone, $IC_{50}B$ is the IC_{50} of the fraction alone, [A] is the IC_{50} of cryptotanshinone in combination with fraction, and [B] is the IC_{50} of fraction in combination with cryptotanshinone. Synergy ≡ $\Sigma FIC < 0.5$, additivity ≡ $0.5 < \Sigma FIC < 1.0$, Indifference ≡ $1.0 < \Sigma FIC < 4.0$, Antagonism ≡ $\Sigma FIC > 4.0$.

Of the eight fractions tested, SM-3 inhibited bacterial growth most strongly, and was prioritized for chromatographic fractionation, yielding 4 simplified fractions (SM-3-1 through SM-3-4). The first fraction, SM-3-1, possessed antimicrobial activity, while the other fractions did not (Appendix C, Figure S24A). We expected that synergists had been separated from cryptotanshinone during the chromatographic separation process and tested the inactive fractions (SM-3-2 through SM-3-4) for synergy. Isobolograms and ΣFIC values for each of these fractions (Appendix C, Figures S24B-S24D) revealed that all three fractions had synergistic activity, with ΣFIC values ranging from 0.14-0.40. Fractions SM-3-2, SM-3-3, and SM-3-4 were chromatographically separated into 21

simplified fractions. Because cryptotanshinone was no longer present at biologically relevant concentrations, it was spiked at sub-lethal concentrations (3 µg/mL) into samples for biological testing so that combination effects could be observed (Figure 33). This approach revealed several fractions with greater than predicted activity (Figure 36A). A subset of fractions was prioritized for synergy testing, revealing fractions that had additive activity and others that had synergistic activity (Table 10).

In the third round of chromatographic separation, fractions were identified that had lower than predicted activity, several of which had sufficient material for biological testing (fractions SM-3-2-7, SM-3-3-2, SM-3-4-1, and SM-3-4-2) (Figure 36A). However, when they were tested for antagonism in a checkerboard assay (127), they had ΣFIC values of 1.25 (SM-3-2-7) or 2.0 (SM-3-3-2, SM-3-4-1, and SM-3-4-2) which we have classified as "noninteractive" (Table 10). There is some inconsistency in the field in determining the ranges for antagonism, and several researchers have considered ΣFIC indices ≥ 2.0 to be indicative of antagonism (319-321). However, we have adopted a more conservative approach, as recommended by Odds (59) and van Vuuren and Viljoen (9), in which antagonistic interactions are defined as having ΣFIC values greater than 4.0. This range takes into account the variability of *in vitro* antimicrobial susceptibility testing, in which a minimum inhibitory concentration can be placed within a three-dilution range (the MIC ± 1 dilution) (59). As such, the more conservative approach enables better interpretation of pharmacological interactions and avoids reproducibility errors when compared to less conservative approaches. It is also important to recognize that interaction between mixtures may differ with different concentrations of compounds/fractions, and

experiments using a single fixed ratio cannot reveal the nature of interactions between mixtures. While the activity index provides a subset of fractions to prioritize for follow up testing, completion of checkerboard assays is critical to define the nature of interactions between samples.

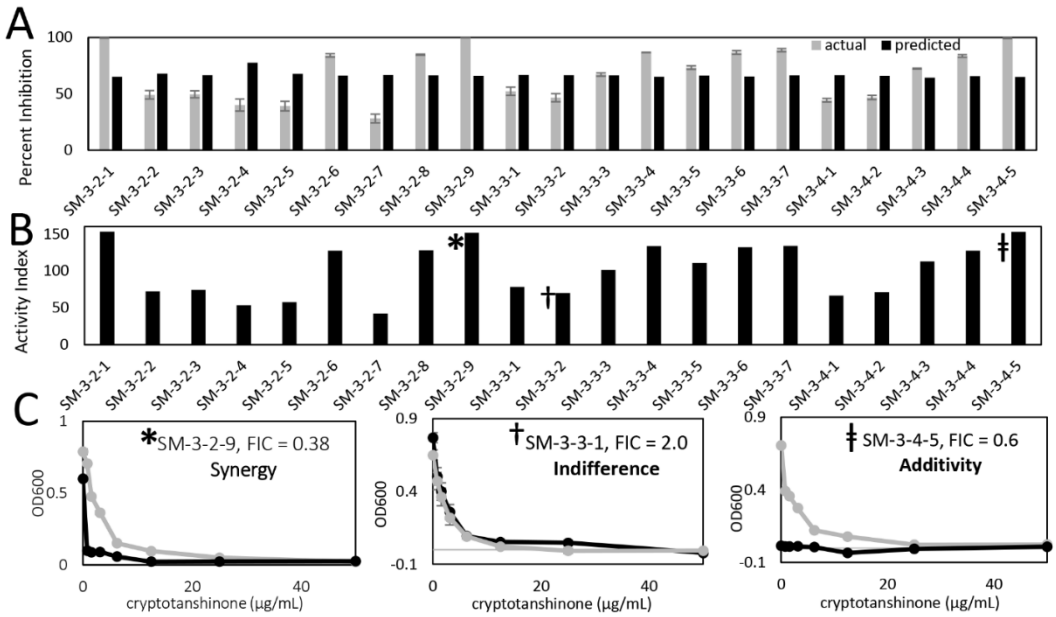


Figure 36. Predicted and Actual Activities of Third Stage Fractions Resulting from Chromatographic Separation of the *Salvia miltiorrhiza* Fractions SM-3-2, SM-3-3, and SM-3-4 (see Fractionation Scheme in Appendix C, Figure S23) where Black Bars Represent the Antimicrobial Activity of Each Fraction due to Cryptotanshinone (Predicted using Peak Area of Cryptotanshinone and Dose Response Curves of Cryptotanshinone Alone) and Gray Bars Represent the Actual Activity of the Fraction at 100 µg/mL. Cryptotanshinone served as a positive control, and its MIC (25 µg/mL) is consistent with previous reports (285). **B.** Activity indices of fractions SM-3-2-1 through SM-3-4-5, where bars represent the extent to which each fraction enhances or suppresses the activity of cryptotanshinone. **C.** Selected dose response curves of cryptotanshinone with (black) and without (gray) 100 µg/mL of synergistic (left), indifferent (middle), and additive (right) fractions. Selected fractions correspond with symbols in panel B.
Activity indices were calculated using equation 1: activity index = actual activity/predicted activity × 100.

Selectivity ratio analysis guided by the activity index predicts compounds contributing to activity and characterizes the nature of their interactions

The first steps of the Simplify approach enabled the prioritization of a subset of *S. miltiorrhiza* fractions whose activity was not explained by the presence of cryptotanshinone alone. While this was helpful for identifying additive and synergistic *mixtures,* it was still unclear which compounds contained in the mixtures were responsible for the observed mismatch between predicted and observed activity. To identify putative active constituents, partial least squares (PLS) analysis was conducted. Rather than use raw biological activity data to guide the analysis, as has been done in previous studies (18, 132, 145), we used the activity index (equation 1) as a measure of the extent to which each fraction enhanced or suppressed the activity of cryptotanshinone.

Using the activity index to guide identification of putative active compounds, two PLS models were produced and visualized with selectivity ratio plots. In these plots, each variable (unique *m/z* – retention time pair) is plotted on the x-axis, and the selectivity ratio is plotted on the y-axis. The selectivity ratio represents the extent to which each variable is associated with biological activity, and is a ratio of the explained to residual variance (149). Because fractions with activity indices > 110 possessed additive or synergistic activity (Figure 36, Table 10), variables possessing high selectivity ratios are most likely be synergists and additives.

Table 10. IC_{50}, MIC, ΣFIC Indices, and Activity Indices (AI) of *S. miltiorrhiza* Extracts in Combination with Cryptotanshinone. IC_{50} and MIC values represent concentrations to inhibit bacterial growth (strain USA300 LAC AH1263) (234) by 50 or 100%, respectively, and represent values of the extract alone, while ΣFIC values indicate the degree of interaction between extracts and cryptotanshinone.

	IC_{50} (μg/mL)*	MIC (μg/mL)	ΣFIC †	AI ‡
SM-1	12.9 ± 1.7	≤ 25	0.38, synergy	--
SM-3	9.8 ± 1.5	≤ 25	0.19, synergy	--
SM-5	11.4 ± 1.5	≤ 25	0.75, additivity	--
SM-3-2	> 100	> 100	0.26, synergy	--
SM-3-3	> 100	> 100	0.40, synergy	--
SM-3-4	46.0 ± 7.2	≤ 100	0.14, synergy	--
SM-3-2-1	> 100	> 100	0.31, synergy	153
SM-3-2-7	> 100	> 100	1.25, indifference	42
SM-3-2-8	> 100	> 100	0.38, synergy	128
SM-3-2-9	> 100	> 100	0.38, synergy	151
SM-3-3-2	> 100	> 100	2.0, indifference	70
SM-3-4-1	> 100	> 100	2.0, indifference	67
SM-3-4-2	> 100	> 100	2.0, indifference	71
SM-3-4-3	> 100	> 100	0.75, additivity	113
SM-3-4-4	> 100	> 100	< 1.0, additivity §	127
SM-3-4-5	12.3 ± 6.3	≤ 25	0.60, additivity	153

* ± standard error
† ΣFICs were calculated using equation 2. Synergy ≡ ΣFIC < 0.5, additivity ≡ 0.5 < ΣFIC < 1.0, Indifference ≡ 1.0 < ΣFIC < 4.0, Antagonism ≡ ΣFIC > 4.0.
‡ Activity indices were calculated for third stage fractions only, which were used to produce SR models.
§ the highest concentration tested was 100 μg/mL, which did not achieve 50% inhibition. To achieve a conservative estimate of activity, however, 100 μg/mL was chosen as the IC_{50} of SM-3-4-4 to calculate the ΣFIC using equation 2 and yielded a result of 1.0. However, since the actual IC_{50} of SM-3-4-4 is higher than 100, the ΣFIC is lower than 1.0, and can be categorized as additive.

The first selectivity ratio model was built using mass spectral data and activity indices from additive and indifferent fractions SM-3-4-1 through SM-3-4-5 (Appendix C, Figure S23) containing 1263 individually detected ions. This internally cross-validated model, used to predict additive compounds, generated 3 components that accounted for 97.1% of the independent (mass spectral) and 89.9% of the dependent (activity indices) variation. The first selectivity ratio plot was generated to visualize ions that corresponded to increased activity indices due to additivity (Figure 37A). Of the 1263 ions included in

the model, only 117 were assigned a selectivity ratio greater than 0. The top ten predicted additives are listed in Table 11.

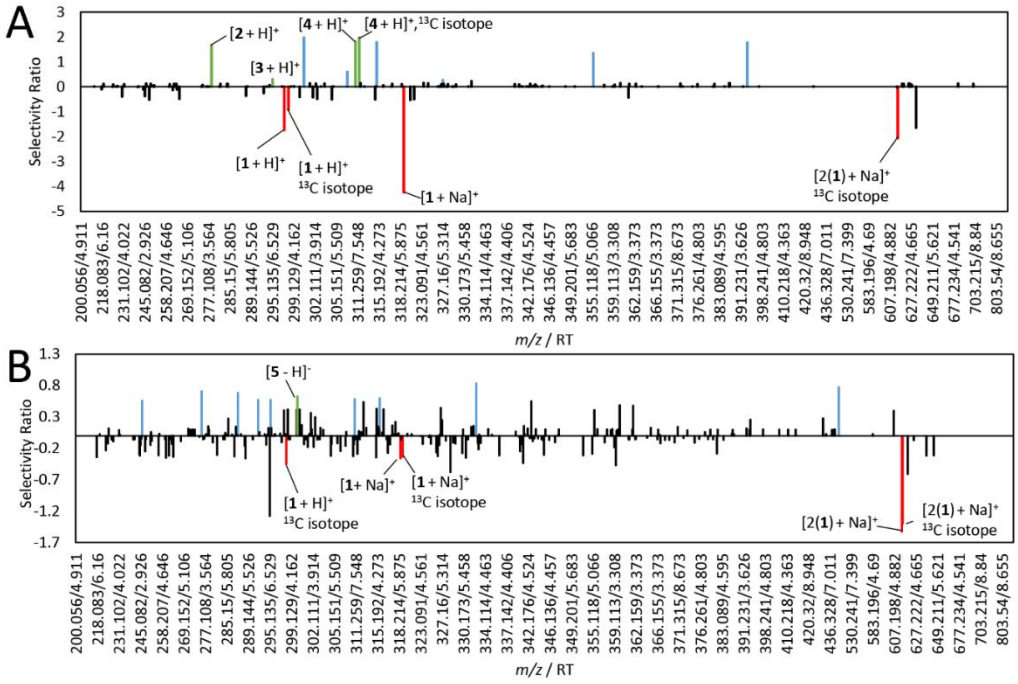


Figure 37. Selectivity Ratio Models Guided by Activity Indices used to Predict Ions Contributing to Additivity and Synergy. Higher selectivity ratios correspond with variables (*m/z* - retention time pairs) that are more likely to contribute to activity. The top ten contributors to activity have been colored green or blue in each model. Importantly, each chemical compound can result in more than one *m/z*-retention time pair because of the numerous isotopes and adducts detected using MS. This provides an additional level of confirmation for the efficacy of analysis, particularly when multiple variables representing a single compound (i.e. compound **4**), are identified as putatively active. Putatively active compounds that have been confirmed by NMR or MS-MS fragmentation patterns have been colored in green, ions corresponding with cryptotanshinone (compound **1**) have been marked in red, and unidentified variables have been marked in blue. Cryptotanshinone (compound **1**) is not correlated with activity in either model because it was spiked in equal concentrations to all fractions under analysis and does not change with changes in bioactivity. **A.** Selectivity ratio plot predicting additive compounds built using data from fractions SM-3-4-1 through SM-3-4-5. Dihydrotanshinone 1 (compound **2**), tanshinone IIA (compound **3**), and 1-oxocryptotanshinone (compound **4**) were identified among the top ten contributors to additive activity. **B.** Selectivity ratio plot predicting synergistic compounds built using data from fractions SM-3-2-1 through SM-3-2-9. Sugiol (compound **5**) was identified as the fifth top contributor to synergistic antimicrobial activity.

The second selectivity ratio model was built using synergistic and indifferent fractions SM-3-2-1 through SM-3-2-9 (Appendix C, Figure S23) using activity indices

and peak area data of 1263 individually detected ions. An internally cross-validated model was produced to predict synergistic compounds and consisted of 3 components utilizing 51.8% of the variability in the mass spectral data to explain 91.4% of activity index variation across fractions. 127 ions were assigned a selectivity ratio greater than 0. This selectivity ratio plot was used to identify the top ten predicted synergists (Figure 37B, Table 11).

Model predictions correctly identified additive and synergistic compounds contributing to the overall antimicrobial activity of *S. miltiorrhiza*

Selectivity ratio models guided by activity indices enabled the prioritization of several compounds likely to possess additive or synergistic activity (Figure 37, Table 11). Two selectivity ratio models were produced, guided by the activity index, enabling the prediction of putative additive compounds in one model (Figure 37A) and synergistic compounds in another (Figure 37B). Using the selectivity ratios of individual constituents, we were able to identify a subset of 20 putatively active compounds from the 1263 ions detected. From the selectivity ratio plots, the dominant marker ions were identified and prioritized for follow up testing. Two of the top ten predicted additives, dihydrotanshinone I (compound **2**) and tanshinone IIA (compound **3**) were identified by comparison of mass spectral fragmentation patterns of standard compounds with compounds detected in *S. miltiorrhiza* fractions (Appendix C, Figures S25 and S26). Purified dihydrotanshinone I and tanshinone IIA were tested in combination with cryptotanshinone as previously described (127, 297) to confirm predictions of additivity.

Table 11. Top Ten Ions Predicted from Both Additive and Synergistic Selectivity Ratio Models. Notably, several of the model predictions were not available as standards and were not present at high enough concentration to isolate and confirm identities. As such, the activity of *S. miltiorrhiza* is likely more complex than represented by the compounds we could identify.

m/z	**Retention time**	**Ionization mode**	**Compound identity**	**Selectivity ratio**	**Predicted activity**
245.117	4.586	+		0.561	Synergistic
275.129	5.438	+		0.71	Synergistic
279.103	5.217	+	Dihydrotanshinone I*	1.66	Additive †
287.164	6.102	+		0.686	Synergistic
292.155	5.456	+		0.574	Synergistic
295.135	6.529	+	Tanshinone IIA *	0.300	Additive †
295.136	5.523	+		0.573	Synergistic
299.201	5.828	-	Sugiol ‡	0.627	Synergistic †
301.085	5.239	+		1.97	Additive
309.113	5.172	+		0.607	Additive
311.129	4.484	+	1-oxocryptotanshinone ‡	1.82	Additive
311.129	4.885	-		0.584	Synergistic
312.131	4.475	+	1-oxocryptotanshinone ‡	1.97	Additive
315.159	3.761	+		1.79	Additive
315.196	5.625	+		0.598	Synergistic
327.160	5.314	+		0.268	Additive
332.185	5.621	+		0.834	Synergistic
355.152	3.793	+		1.35	Additive
393.287	4.809	+		1.78	Additive
460.196	5.459	+		0.776	Synergistic

* compound identity confirmed by comparing MS-MS patterns of a pure standard
† activity confirmed by running full checkerboard assays
‡ compound identity confirmed by NMR

Dihydrotanshinone I and tanshinone IIA had ΣFIC values of 0.68 and 0.61, respectively, confirming the predictions from the selectivity ratio analysis (Table 12). They were also each antimicrobial in isolation, with MIC values $\leq$ 6.25 and 25 μg/mL and IC_{50} values of 2.2 ± 0.4 and 15.0 ± 8.4 (for dihydrotanshinone I and tanshinone IIA, respectively). An additional predicted additive compound, with an $[M+H]^+$ of 311.1277, representing the top contributor to activity, was prioritized for isolation. This compound, 1-oxocryptotanshinone (compound **4**) was isolated following 2 stages of normal-phase flash chromatography and 2 stages of reversed-phase chromatography. This compound has not previously been isolated from *S. miltiorrhiza.* Unfortunately, compound **4** was not present

in sufficient quantity for additivity predictions to be confirmed. However, given the structural similarity of compound **1** and compound **4**, it is likely that compound **4** contributes to the overall activity of the extract. Dose-response curves for all tested compounds are provided as supporting information (Appendix C, Figure S27).

Table 12. IC_{50}, MIC, and ΣFICs of Pure Compounds from *S. miltiorrhiza* in Combination with Cryptotanshinone. IC_{50} and MIC values represent single compound concentrations to inhibit bacterial growth (strain USA300 LAC AH1263) (234) by 50 or 100%, respectively, while ΣFIC values indicate the interactions between pure compounds and cryptotanshinone. IC_{50} values were calculated using a 4-parameter logistic curve.

	IC_{50} (μg/mL) *	**MIC (μg/mL)**	**ΣFIC** †
Cryptotanshinone	5.9 ± 2.2	≤ 25	--
Dihydrotanshinone I	2.2 ± 0.4	≤ 6.25	0.68, additivity
Tanshinone IIA	15.0 ± 8.4	≤ 25	0.61, additivity
Sugiol	> 100	> 100	0.28, synergy

* ± standard error

† ΣFICs were calculated using equation 2.

Numerous compounds were identified as potentially contributing to the synergistic activity of *S. miltiorrhiza* fractions. One predicted synergist, sugiol, was isolated using a combination of normal- and reversed-phase chromatography and its structure confirmed by NMR (Appendix C, Figures S28 – S33, Appendix B, Table S8) (322). Sugiol (compound 5), the 5th top contributor to synergy according to model predictions (Figure 37), was tested in combination with cryptotanshinone to confirm synergistic activity. Indeed, sugiol possessed an ΣFIC value of 0.28 in combination with cryptotanshinone, confirming its activity as a synergist. Alone, sugiol did not possess antimicrobial activity, with IC_{50} and MIC values >100 μg/mL. Despite its lack of activity in isolation, when combined with cryptotanshinone, sugiol induced nearly a four-fold drop in cryptotanshinone's IC_{50}, lowering it from from 5.89 to 1.56 μg/mL, (Figure 38).

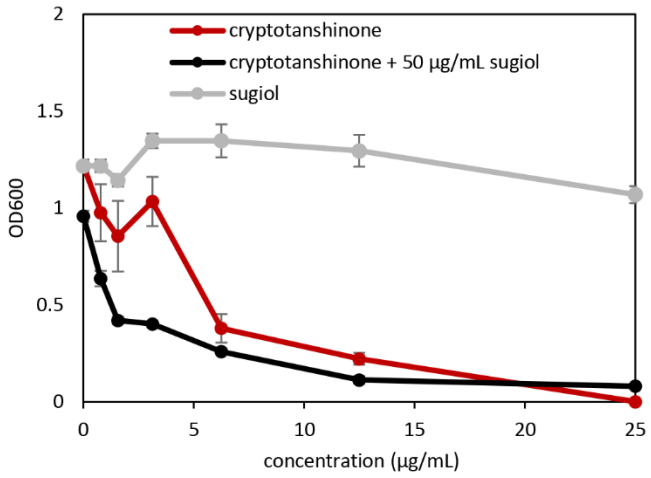


Figure 38. Dose-Response Curves of Cryptotanshinone Alone, Cryptotanshinone in Combination with Sugiol (Fixed Concentration of 50 µg/mL Sugiol), and Sugiol Alone. Error bars represent standard error (not visible for some data points because they are smaller than the point size). Notably, sugiol did not have any antimicrobial effect when tested individually at concentrations ≤ 100 µg/mL. However, in combination with cryptotanshinone, it causes a four-fold drop in cryptotanshinone's IC_{50} (increased potency), illustrating that sugiol has synergistic effects.

Discussion

The research described herein represents the first example of a bioinformatics approach being used to predict compounds contributing to combination effects within complex mixtures and to characterize the nature of their interactions prior to their isolation. Using *S. miltiorrhiza* as a model organism, the Simplify approach enabled the identification and differentiation of three additive compounds and one synergistic compound. The Simplify approach is compatible with existing bioassay-directed fractionation workflows and minimizes the time required to go from the discovery of combination effects within a mixture to lead molecules contributing to synergy and additivity. While we did not observe antagonist effects in this study, it follows that the approach described herein could also be utilized to identify constituents that mask

biological activity in cases where a suppression in activity was observed as a result of antagonism.

Using the second selectivity ratio plot designed to predict synergists (Figure 37B), sugiol (compound **5**) was correctly predicted to synergize with cryptotanshinone. Although this compound has been identified from *S. miltiorrhiza* previously (323), this is the first report of its ability to synergize the antimicrobial activity of cryptotanshinone. Using traditional bioassay-guided fractionation approaches, this compound would have been missed due to its inactivity in isolation. However, using the Simplify approach, this compound was not only identified as active, but its activity as a synergist was predicted prior to its isolation.

Interestingly, the fractions used to produce the selectivity ratio plot for predicting additive compounds (Figure 37A), which all possessed additive rather than synergistic activity, came from the separation of fraction SM-3-4, which itself was synergistic. This observation illustrates the complexity of botanical samples and the interactions present within them. It is possible that synergistic compounds were lost during the chromatographic separation process due to irreversible binding to the column, or that multiple constituents were required for the observed synergistic effect of fraction SM-3-4. Thus, while the results we provide herein give a more comprehensive picture of the constituents that contribute to the activity of the *S. miltiorrhiza* mixture than was possible using other approaches, it is still only part of the story. Furthermore, due to material limitations, it was not possible to isolate the additional compounds predicted to possess additive or synergistic activity in Figure 37B, again pointing to the possibility of more

compounds contributing to the activity of *S. miltiorrhiza*. This demonstrates an inherent limitation to the Simplify approach; while it may be possible to identify multiple putatively active constituents, material limitations remain a central challenge in natural product discovery efforts. However, the advantage of the Simplify approach is that isolation efforts were guided towards compounds likely to be active (from among the more than 1200 ions detected). Notably, several predictions generated with this approach were validated; all compounds tested based on the predictions possessed the expected antimicrobial or synergistic activity.

Increasing numbers of studies demonstrate that individual mixture constituents often behave very differently in isolation than they do within a complex mixture (18, 144, 324-326). Environmental exposures, for example, occur as complex mixtures which may include additive and synergistic effects (324). Too often, it is assumed that the behavior of a mixture can be described by the presence of just a few known constituents. Indeed, each biological system relies on diverse chemical interactions; not only do organisms themselves represent complex mixtures, but they interact with a vast array of organic and inorganic chemicals for survival. Currently, there is a gap in the way that we understand complex mixtures, whether they be natural product extracts, environmental contaminants, or human microbiota, because these mixtures do not exert biological effects equal to the sum of their individual constituents. With this study, we illustrate the ability of Simplify to provide a more comprehensive picture of combination effects present within complex mixtures than is possible with more reductionist approaches. We expect that this tool will

be applicable within and outside of the field of natural products drug discovery to illuminate chemical interactions that occur within complex mixtures.

Methods

General experimental procedures

UPLC-MS analysis was conducted using a Thermo-Fisher Q-Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific, MA, USA) coupled to an Acquity UPLC system (Waters Corporation, Milford, MA, USA). UPLC separations were achieved using a reversed-phase column (BEH C18, 1.7 µm, 2.1 × 50 mm, Waters Corporation, Milford, MA, USA). Each sample was analyzed in triplicate at a concentration of 0.1 mg/mL in methanol with a 3µL injection. A gradient of water (solvent A) and acetonitrile (solvent B), each containing 0.1% formic acid, was employed. The gradient began at 90:10 (A:B) and was held for 0.5 min. From 0.5-8.0 min, the ratio was increased to 0:100 (A:B) and held until 8.5 min. From 8.5-9.0 min, the starting conditions were re-established, after which the gradient was held at 90:10 (A:B) until 10.0 min. Mass analysis was conducted in both positive and negative ion modes with scan range of 150-1500, capillary temperature of 256°C, S-lens RF level of 50.00, spray voltage of 3.50 kV, sheath gas flow 47.50, and auxiliary gas flow of 11.25. A data-dependent method was used in which the four ions with the highest signal intensity within each scan were fragmented using an HCD of 35.0.

Plant material and extraction

Fresh roots of *Salvia miltiorrhiza* were collected on November 8, 2016 at the Chicago Ashram (Batch #CRS1016F1, 41W501 Keslinger Rd, Elburn, IL 60119). Plant

identification was conducted by Richard A. Cech at Strictly Medicinal Seeds, and a voucher specimen grown from the same seed line was collected September 3, 2017 and deposited at the herbarium of the University of North Carolina at Chapel Hill (NCU652634).

Fresh *S. miltiorrhiza* roots were dug, washed, and chopped fresh and airdried, yielding 500 g of dried material. The dried roots were ground using a Wiley Mill Standard Model No. 3 (Arthur Thomas Company) and extracted in MeOH at 160 g/L for 24 hours prior to filtering. This process was repeated with the same root material every 24 hours for 72 hours. The final MeOH extract was concentrated *in vacuo* and the residue was partitioned between 10% aqueous MeOH and hexane (1:1) for defatting. The aqueous MeOH layer was then partitioned using 4:5:1 EtOAc:MeOH:H_2O. Finally, the EtOAc layer was washed with a 1% NaCl solution (1:1) to remove hydrosoluble tannins. The resulting EtOAc extract was dried under nitrogen yielding 18.32 g of material.

Chromatographic separation and isolation

The isolation scheme is provided as Supporting Information (Appendix C, Figure S23). See SI Appendix A, Supplementary Protocols (Protocol S3) for details on chromatographic separation of complex fractions, and isolation of compounds **1** (cryptotanshinone), **4** (1-oxocryptotanshinone), and **5** (sugiol).

*Sugiol (**5**):* white amphorphous powder; HRESIMS *m/z* 301.2159 $[M+H]^+$ (calculated for $C_{20}H_{29}O_2^+$, 301.2167, -2.6 ppm). MS/MS data of the isolated compound match fragmentation patterns of the compound found within *S. miltiorrhiza* extract (Appendix C, Figure S28). NMR data from the literature are inconsistent and incomplete

(322, 327-332), and improved spectra are provided as Supporting Information (Appendix B, Table S8, Appendix C, Figures S29-S33). ^{1}H NMR (500 MHz, $CDCl_3$) δ 0.92 (3H, s, H-18), 0.98 (3H, s, H-19), 1.21 (3H, s, H-20), 1.24 (3H, d, J=6.9 Hz, H=16), 1.25 (1H, m, H-3α), 1.26 (3H, d, J=6.9 Hz, H-17), 1.53 (1H, m, H-1α), 1.53 (1H, m, H-3β), 1.67 (1H, m, H-2α), 1.76 (1H, tt, J=13.6, 3.3 Hz, H-2β), 1.85 (1H, dd, J=13.7, 4.0 Hz, H-5), 2.23 (1H, dt, J=11.9, 2.8 Hz, H-1β), 2.58 (1H, dd, J=18.1, 13.8 Hz, H-6β), 2.68 (1H, dd, J=18.1, 4.0 Hz, H-6α), 3.12 (1H, hept, J=6.9 Hz, H-15), 6.68 (1H, s, H-11), 7.90 (1H, s, H-14) (Appendix C, Figure S29). ^{13}C NMR (125 MHz, $CDCl_3$) δ 18.97 (CH_2, C-2), 21.45 (CH_3, C-19), 22.42 (CH_3, C17), 22.55 (CH_3, C-16), 23.33 (CH_3, C-20), 26.88 (CH, C-15), 32.65 (CH_3, C-18), 33.36 (C, C-4), 36.13 (CH_2, C-6), 37.95 (C, C-10), 37.97 (CH_2, C-1), 41.42 (CH_2, C-3), 49.53 (CH, C-5), 110.03 (CH, C-11), 124.78 (C, C-8), 126.63 (CH, C-14), 132.63 (C, C-13), 156.52 (C, C-9), 158.15 (C-OH, C-12), 198.68 (ketone, C-7) (Appendix C, Figure S30). HSQC experiments (500 MHz, $CDCl_3$) were used to assign overlapping peaks (Appendix C, Figure S31). HMBC (500 MHz, $CDCl_3$) data are also provided (Appendix C, Figure S32). A previous report was conducted in DMSO-d_6 (322), and for further confirmation on the identity of this compound, we ran an additional ^{1}H NMR (500 MHz, DMSO-d_6), which matched literature values (Appendix C, Figure S33) (322).

*Cryptotanshinone (**1**):* red crystalline solid; HRESIMS *m/z* 297.1487 $[M+H]^+$ (calculated for $C_{19}H_{21}O_3^+$, 297.1490, 1.01 ppm); ^{1}H NMR (500 MHz, $CDCl_3$) and ^{13}C NMR (125 MHz, $CDCl_3$) are consistent with previous reports (333), and are provided as

supporting information (Appendix C, Figures S35 and S35), and MS/MS data of isolated cryptotanshinone match those of the purchased standard (Appendix C, Figure S36).

*1-oxo-cryptotanshinone (**4**):* orange amorphous powder; HRESIMS *m/z* 311.1277 $[M+H]^+$ (calculated for $C_{19}H_{19}O_4^+$, 1.93 ppm); ^1H NMR (500 MHz, $CDCl_3$) are consistent with previous reports (334), and is provided as supporting information (Appendix C, Figure S37). Compound degradation occurred before additional spectra could be obtained.

Antimicrobial assays

To evaluate antimicrobial activity, a broth microdilution assay was conducted on each sample using a clinically relevant strain of methicillin-resistant *Staphylococcus aureus* (strain USA300 LAC AH1263) (234). Cells, diluted to 1.0×10^5 colony forming units (CFU) per milliliter calculated using absorbance at 600 nm (OD_{600}) values with a Synergy H1 microplate reader (Biotek, Winooski, VT, USA), were inoculated into Müeller-Hinton broth (MHB) with or without *S. miltiorrhiza* extract or purified compounds. Assays were completed in triplicate using standard protocols from the Clinical Laboratory Standards Institute (CLSI) (236). Stock solutions of pure compounds and complex extracts were prepared in DMSO and diluted with MHB for a final concentration in test wells of 2% DMSO and a sample concentration ranging from 0-100 µg/mL. Full dose-response curves of cryptotanshinone (compound **1**) were conducted during each screening, and served as positive control (313, 314, 317). Minimum inhibitory concentrations (MIC) were calculated for each pure compound, defined as the concentration at which no statistically significant differences in OD_{600} were found

between wells containing samples but no bacteria and the treated samples. Dose-response curves were produced using SigmaPlot (v.13, Systat Software, San Jose, CA, USA) and plotted with a four-parameter logistic model.

Activity prediction and production of activity index

To identify a subset of *S. miltiorrhiza* fractions for synergy testing, the Simplify workflow was employed (Figure 33). *S. miltiorrhiza* extract and resulting chromatographic fractions were subjected to UPLC-MS as described in *General experimental procedures.* In tandem, an external calibration curve of cryptotanshinone (ranging from 0-50 µg/mL) was produced (Appendix C, Figure S38). Using the linear range of the calibration curve, concentrations of cryptotanshinone within each sample were calculated. After two rounds of fractionation, synergists had been separated from cryptotanshinone during chromatographic separation. To avoid mis-identifying these fractions as inactive, a sub-lethal concentration of cryptotanshinone (3 µg/mL) was spiked into each sample prior to both antimicrobial activity assessment and liquid chromatography - mass spectrometry analysis. Antimicrobial dose-response curves of cryptotanshinone were used to predict activity of fractions based on their concentration of cryptotanshinone. The predicted and observed biological activities of fractions were compared, and those illustrating a mismatch in activity greater than 10% were prioritized for checkerboard assays to disentangle combination effects provided that they also contained sufficient material for follow up testing.

Synergy assessment

Broth microdilution checkerboard assays (127) were utilized to pinpoint the type of combination effects present in *S. miltiorrhiza* fractions and their impact on the biological activity of cryptotanshinone. A subset of *S. miltiorrhiza* fractions (Table 10) and pure compounds (compounds **2, 3,** and **5**) were tested in combination with cryptotanshinone, with fraction concentrations ranging from 0-100 µg/mL and cryptotanshinone ranging from 0-25 µg/mL. The vehicle control consisted of 2% DMSO in MHB. Fractional inhibitory concentration indices (ΣFICs) were calculated using equation 2. Because the IC_{50} values often fall between tested concentrations, the IC_{50} values for A and B in combination were identified as the lowest tested concentrations that led to ≥ 50% growth inhibition. Conservative values were chosen to assign combination effects to avoid data misinterpretation, as recommended by van Vuuren and Viljoen (9). Synergistic effects are defined as interactions having ΣFIC ≤ 0.5, additive effects range from 0.5 to 1.0, non-interactive effects range from 1.0 to 4.0, and antagonistic effects have an ΣFIC ≥ 4.0 (9).

Metabolomics data analysis

Baseline correction/MZmine parameters

Datasets acquired in positive and negative modes were analyzed, aligned, and filtered using MZMine 2.21.2 (http://mzmine.sourceforge.net/) (239). Raw data files of each sample, as well as its triplicate injections, were uploaded into MZMine for peak picking. Chromatograms were built for all *m/z* values that were detected for longer than 0.1 min. Modeling was completed with the third set of fractions (See fractionation

scheme, Appendix C, Figure S23) using a noise level (absolute value) of 7.0×10^6 for negative mode and 2.0×10^6 for positive mode. The *m/z* tolerance was 5ppm (or 0.0001 Da), and the intensity variation tolerance was 20%. Peaks were aligned into a single peak if they eluted within 0.2 min from one another and had less than 5 ppm difference in *m/z* values. Data consisting of retention time, *m/z* values, and peak area for both negative and positive ions was imported into Excel (Microsoft, Redmond, WA, USA) and combined as a single peak list. Dataset reduction was completed in Excel before statistical analysis. First, all *m/z*- retention time pairs that had peak areas above 1.0×10^7 in the methanol blank were removed from analysis. Compounds containing an *m/z* ratio below 200 or above 900 were also removed, as were compounds eluting before 1.5 or after 9 min. Following dataset reduction, biological data were added in the form of activity indices calculated using percent inhibition data at 100 µg/mL. The results of the checkerboard assays of the prioritized subset of fractions were used to divide samples into two groups. One group had high activity indices due to synergy and the other group consisted of fractions that had high activity indices due to additivity. Fractions that did not have a mismatch, or had a lower than predicted activity, were included in both models after checkerboard assays revealed them to be non-interactive. Data matrices for these two sample subsets were imported into Sirius version 10.0 (Pattern Recognition Systems AS, Bergen, Norway) (240) for statistical analysis.

Selectivity ratio analysis

Sirius version 10.0 statistical software (240) was used to filter contaminants and to generate selectivity ratio models to predict additives and synergists in *S. miltiorrhiza*

samples. Prior to selectivity ratio analysis, triplicate injections of samples were subjected to hierarchical cluster analysis and filtering of interferents as described in a recent publication (276). If triplicate injections did not cluster, spectral variables were inspected, and variables illustrating high peak area variability (above 1.5×10^8) within triplicate injections were removed, as were their associated in-source fragments, mass spectral adducts, and isotopes (Appendix B, Table S8). Following the removal of chemical interferents, two selectivity ratio models were produced, one using a subset of synergistic and indifferent fractions (SM-3-2-1 through SM-3-2-9), and one using a subset of additive and indifferent fractions (SM-3-4-1 through SM-3-4-5). Each subset underwent an internally cross-validated PLS analysis using 100 iterations and a significance level of 0.05. Activity indices at the 100 µg/mL level were used as the dependent variable guiding separation between groups. Internal algorithms of the Sirius program were computed, resulting in selectivity ratio plots that identified putative additive compounds (additivity model) and synergistic compounds (synergy model). To simplify model interpretation and remove possible correlated noise from model datasets, a final filtering step, in which variables showing less than 10% peak area variance across samples were given a selectivity ratio of 0, was conducted (144).

Acknowledgements

This research was supported by the National Center for Complementary and Integrative Heath of the National Institutes of health under grant numbers 5 T32 AT008938 and 1 R15 AT010191. Additionally, we acknowledge Richard A (Richo) Cech for his provision of plant material, Dr. Alexander Horswill for providing microbial

strains, and Dr. Olav Kvalheim for his consultation in data analysis. Sonja Knowles, Dr. Nicholas Oberlies, and Dr. Ron Venters provided technical assistance with NMR analysis. Mass spectrometry analyses were conducted in the Triad Mass Spectrometry Facility at the University of North Carolina at Greensboro with assistance from Dr. Daniel A. Todd.

CHAPTER VII

CONCLUDING REMARKS

The studies described here demonstrate the efficacy of mass spectrometry-based tools for understanding the vast chemical landscape of botanical medicines. These tools can aid in the identification of active constituents within complex mixtures by integrating chemical profiles with biological activity profiles, enabling the targeting of active, rather than abundant, constituents. Chapter III described the utilization of bioassay-guided fractionation, biochemometric analysis, and molecular networking to predict putative active constituents from *Angelica keiskei*. These predictions were confirmed, and a new activity for a compound previously believed to be inactive was discovered.

Chapter IV described the development of a data filtering tool using hierarchical cluster analysis (HCA) of technical replicates to remove chemical interferents from metabolomics datasets. Using this process, 128 contaminant ions were identified that likely originated from the UPLC-MS system, enabling improved metabolomics analysis and highlighting the importance of technical replicates for metabolomics studies. These results also challenged the assumption that contamination is consistent across samples. The project outlined in Chapter V took a closer look at how data acquisition and data processing parameters affect biochemometric analysis by using a mixture of known composition and comparing numerous selectivity ratio models subjected to a variety of data processing and data acquisition techniques. This project highlighted the variety of

biological, chemical, and analytical factors that can complicate metabolomics analysis and provides guidelines for future studies.

Chapter VI outlines the development and application of a novel approach, Simplify, which identifies the extent to which a given mixture enhances or reduces the activity of a known antimicrobial. Using this information, the Simplify approach builds selectivity ratio plots capable not only of predicting directly active components, but also enables the identification of components that contribute indirectly to activity through synergistic and antagonistic mechanisms. This is the first documented example in which synergistic compounds have been predicted as synergists prior to their isolation and illustrates the efficacy of this approach for understanding how mixtures work in concert. This approach is expected to be valuable beyond the field of natural products, applicable to any field aiming to identify how complex mixtures work in concert and is expected to serve as a launching point for the comprehensive evaluation of mixtures in future studies.

Overall, these studies, combining microbiology, metabolomics analysis, and analytical chemistry, emphasize the advantages that mass spectrometry provides for understanding medicinal natural products. The volume of data that mass spectrometry provides is both a blessing and a curse, and care must be taken, particularly when processing and interpreting models built on these massive datasets, to extract meaningful information and identify the biological patterns contained within them. Based upon this work, continued improvement of biological measurements, data filtering protocols, and development and interpretation of multivariate models is likely to provide an important avenue for natural product drug discovery.

REFERENCES

1. Petrovska BB (2012) Historical review of medicinal plants' usage. *Pharmacogn Rev.* 6(11):1-5.

2. Kelly K (2009) *The history of medicine* (Facts on file).

3. Bandaranayake WM (2006) Quality Control, Screening, Toxicity, and Regulation of Herbal Drugs. *Modern Phytomedicine: Turning Medicinal Plants into Drugs*, eds Ahmad I, Aqil F, & Owais M (WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany), pp 25-57.

4. Bodeker G & Ong C-K (2005) *WHO global atlas of traditional, complementary and alternative medicine* (World Health Organization).

5. Ekor M (2014) The growing use of herbal medicines: issues relating to adverse reactions and challenges in monitoring safety. *Front Pharmacol.* 4:177.

6. Black LI, Clarke TC, Barnes PM, Stussman BJ, & Nahin RL (2015) Use of complementary health approaches among children aged 4–17 years in the United States: National Health Interview Survey, 2007–2012. *Natl Health Stat Rep.* (78):1.

7. Clarke TC, Black LI, Stussman BJ, Barnes PM, & Nahin RL (2015) Trends in the use of complementary health approaches among adults: United States, 2002–2012. *Natl Health Stat Rep.* (79):1.

8. Efferth T & Koch E (2011) Complex interactions between phytochemicals. The multi-target therapeutic concept of phytotherapy. *Curr Drug Targets.* 12(1):122-132.

9. van Vuuren S & Viljoen A (2011) Plant-based antimicrobial studies–methods and approaches to study the interaction between natural products. *Planta Med.* 77(11):1168-1182.

10. Wagner H & Ulrich-Merzenich G (2009) Synergy research: approaching a new generation of phytopharmaceuticals. *Phytomedicine.* 16(2-3):97-110.

11. Burfield T & Reekie S-L (2005) Mosquitoes, malaria and essential oils. *Int J Aromather.* 15(1):30-41.

12. Raskin I & Ripoll C (2004) Can an apple a day keep the doctor away? *Curr Pharm Des.* 10(27):3419-3429.

13. Enke CG & Nagels LJ (2011) Undetected components in natural mixtures: how many? What concentrations? Do they account for chemical noise? What is needed to detect them? *Anal Chem.* 83(7):2539-2546.

14. Junio HA, *et al.* (2011) Synergy-directed fractionation of botanical medicines: a case study with goldenseal (Hydrastis canadensis). *J Nat Prod.* 74(7):1621-1629.

15. Stermitz FR, Lorenz P, Tawara JN, Zenewicz LA, & Lewis K (2000) Synergy in a medicinal plant: antimicrobial action of berberine potentiated by 5′-methoxyhydnocarpin, a multidrug pump inhibitor. *Proc Natl Acad Sci.* 97(4):1433-1437.

16. Stermitz FR, Scriven LN, Tegos G, & Lewis K (2002) Two flavonols from Artemisa annua which potentiate the activity of berberine and norfloxacin against a resistant strain of Staphylococcus aureus. *Planta Med.* 68(12):1140-1141.

17. Ulrich-Merzenich G, Panek D, Zeitler H, Vetter H, & Wagner H (2010) Drug development from natural products: exploiting synergistic effects. *Indian J Exp Biol.* 48(3):208-219.

18. Britton ER, Kellogg JJ, Kvalheim OM, & Cech NB (2017) Biochemometrics to identify synergists and additives from botanical medicines: a case study with Hydrastis canadensis (goldenseal). *J Nat Prod.* 81(3):484-493.

19. Bunterngsook B, Eurwilaichitr L, Thamchaipenet A, & Champreda V (2015) Binding characteristics and synergistic effects of bacterial expansins on cellulosic and hemicellulosic substrates. *Bioresour Tech.* 176:129-135.

20. Chevereau G & Bollenbach T (2015) Systematic discovery of drug interaction mechanisms. *Mol Syst Biol.* 11(4):807.

21. Piggott JJ, Townsend CR, & Matthaei CD (2015) Reconceptualizing synergism and antagonism among multiple stressors. *Ecol Evol.* 5(7):1538-1547.

22. Tang J, Wennerberg K, & Aittokallio T (2015) What is synergy? The Saariselkä agreement revisited. *Front Pharmacol.* 6:181.

23. Berenbaum MC (1989) What is synergy. *Pharmacol Rev.* 41:93-141.

24. Rather MA, Bhat BA, & Qurishi MA (2013) Multicomponent phytotherapeutic approach gaining momentum: Is the “one drug to fit all” model breaking down? *Phytomedicine.* 21(1):1-14.

25. Williamson EM (2001) Synergy and other interactions in phytomedicines. *Phytomedicine.* 8(5):401-409.

26. Medina-Franco JL, Giulianotti MA, Welmaker GS, & Houghten RA (2013) Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discov Today.* 18(9-10):495-501.

27. Zimmermann GR, Lehar J, & Keith CT (2007) Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug Discov Today.* 12(1-2):34-42.

28. Borisy AA, *et al.* (2003) Systematic discovery of multicomponent therapeutics. *Proc Natl Acad Sci.* 100(13):7977-7982.

29. Bonapace CR, White RL, Friedrich LV, & Bosso JA (2000) Evaluation of antibiotic synergy against Acinetobacter baumannii: a comparison with Etest, time-kill, and checkerboard methods. *Diagn Microbiol Infect Dis.* 38(1):43-50.

30. Lewis R, Diekema D, Messer S, Pfaller M, & Klepser M (2002) Comparison of Etest, chequerboard dilution and time–kill studies for the detection of synergy or antagonism between antifungal agents tested against Candida species. *J Antimicrob Chemother.* 49(2):345-351.

31. White RL, Burgess DS, Manduru M, & Bosso JA (1996) Comparison of three different in vitro methods of detecting synergy: time-kill, checkerboard, and E test. *Antimicrob Agents Chemother.* 40(8):1914-1918.

32. Wu M & Woolson RF (1998) A comparison of several tests for drug synergy. *Commun Stat Simul Comput.* 27(2):303-327.

33. Shikov AN, Pozharitskaya ON, & Makarov VG (2018) Challenges in the investigation of combinatory modes of action of nutrients and pharmaceuticals. *Synergy.* 7:36-38.

34. Ocana A, Amir E, Yeung C, Seruga B, & Tannock IF (2012) How valid are claims for synergy in published clinical studies? *Ann Oncol.* 23(8):2161-2166.

35. Tam VH, Schilling AN, & Nikolaou M (2005) Modelling time–kill studies to discern the pharmacodynamics of meropenem. *J Antimicrob Chemother.* 55(5):699-706.

36. An J, Zuo G, Hao X, Wang G, & Li Z (2011) Antibacterial and synergy of a flavanonol rhamnoside with antibiotics against clinical isolates of methicillin-resistant Staphylococcus aureus (MRSA). *Phytomedicine* 18(11):990-993.

37. Lederer S, Dijkstra TM, & Heskes T (2018) Additive Dose Response Models: Defining Synergy. *bioRxiv*:480608.

38. Yadav B, Wennerberg K, Aittokallio T, & Tang J (2015) Searching for drug synergy in complex dose–response landscapes using an interaction potency model. *Comp Struct Biotechnol J.* 13:504-513.

39. Chou T-C (2006) Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacol Rev.* 58(3):621-681.

40. Greco WR, Bravo G, & Parsons JC (1995) The search for synergy: a critical review from a response surface perspective. *Pharmacol Rev.* 47(2):331-385.

41. Lee S-i (2010) Drug interaction: focusing on response surface models. *Korean J Anesthesiol.* 58(5):421-434.

42. Zhao L, Au JL-S, & Wientjes MG (2010) Comparison of methods for evaluating drug-drug interaction. *Front Biosci.* 2:241.

43. Bliss CI (1939) The toxicity of poisons applied jointly. *Ann Appl Biol.* 26:858-615.

44. Loewe S (1953) The problem of synergism and antagonism of combined drugs. *Arzneimittelforschung.* 3:285-290.

45. Geary N (2012) Understanding synergy. *Am J Physiol Endocrinol Metabol.* 304(3):E237-E253.

46. Lederer S, Dijkstra TM, & Heskes T (2018) Additive Dose Response Models: Explicit Formulation and the Loewe Additivity Consistency Condition. *Front Pharmacol.* 9:31.

47. Tallarida RJ (2006) An overview of drug combination analysis with isobolograms. *J Pharm Exp Ther.* 319(1):1-7.

48. Cokol M, *et al.* (2011) Systematic exploration of synergistic drug pairs. *Mol Syst Biol.* 7(1):544.

49. Mathews Griner LA, *et al.* (2014) High-throughput combinatorial screening identifies drugs that cooperate with ibrutinib to kill activated B-cell–like diffuse large B-cell lymphoma cells. *Proc Natl Acad Sci.* 111(6):2349-2354.

50. Chou T-C (2010) Drug combination studies and their synergy quantification using the Chou-Talalay method. *Cancer Res.*:0008-5472. CAN-0009-1947.

51. Fitzgerald JB, Schoeberl B, Nielsen UB, & Sorger PK (2006) Systems biology and combination therapy in the quest for clinical efficacy. *Nat Chem Biol.* 2(9):458.

52. Kong M & Lee JJ (2006) A generalized response surface model with varying relative potency for assessing drug interaction. *Biometrics.* 62(4):986-995.

53. Lee JJ, Kong M, Ayers GD, & Lotan R (2007) Interaction index and different methods for determining drug interaction in combination therapy. *J Biopharm Stat.* 17(3):461-480.

54. Zhao W, *et al.* (2014) A new bliss independence model to analyze drug combination data. *J Biolmol Screen.* 19(5):817-821.

55. Greco W, *et al.* (1992) Consensus on concepts and terminology for combined-action assessment: the Saariselkä agreement. *Arch Complex Environ Stud.* 4(3):65-69.

56. Russ D & Kishony R (2018) Additivity of inhibitory effects in multidrug combinations. *Nat Microbiol.* 3(12):1339.

57. Chandrasekaran K, *et al.* (2002) Bilobalide, a component of the Ginkgo biloba extract (EGb 761), protects against neuronal death in global brain ischemia and in glutamate-induced excitotoxicity. *Cell Mol Biol.* 48(6):663-669.

58. Chou T-C & Talalay P (1983) Analysis of combined drug effects: a new look at a very old problem. *Trends Pharmacol Sci.* 4:450-454.

59. Odds FC (2003) Synergy, antagonism, and what the chequerboard puts between them. *j Antimicrob Chemother.* 52(1):1-1.

60. Di Veroli GY, *et al.* (2016) Combenefit: an interactive platform for the analysis and visualization of drug combinations. *Bioinformatics* 32(18):2866-2868.

61. Gilbert B & Alves L (2003) Synergy in plant medicines. *Curr Med Chem.* 10(1):13-20.

62. Rasoanaivo P, Wright CW, Willcox ML, & Gilbert B (2011) Whole plant extracts versus single compounds for the treatment of malaria: synergy and positive interactions. *Malar J.* 10 Suppl 1:S4.

63. Beyerstein BL (2001) Alternative medicine and common errors of reasoning. *Academ Med.* 76(3):230-237.

64. Kaptchuk TJ (2002) The placebo effect in alternative medicine: can the performance of a healing ritual have clinical significance? *Ann Int Med.* 136(11):817-825.

65. Lewith GT, Hyland ME, & Shaw S (2002) Do attitudes toward and beliefs about complementary medicine affect treatment outcomes? *Am J Pub Health.* 92(10):1604-1606.

66. Tausk FA (1998) Alternative medicine: is it all in your mind? *Arch Derm.* 134(11):1422-1425.

67. Ma X, *et al.* (2009) Synergistic therapeutic actions of herbal ingredients and their mechanisms from molecular interaction and network perspectives. *Drug Discov Today.* 14(11-12):579-588.

68. Efferth T, *et al.* (2015) Nobel Prize for artemisinin brings phytotherapy into the spotlight. *Phytomedicine.* 22(13):1-4.

69. Tu YY, Ni MY, Zhong Y, & Li LN (1981) Studies on the constituents of Artemisia annua L. and derivatives of artemisinin. *Zhongguo Zhong Yao Za Zhi.* 6(31).

70. White NJ (2008) Qinghaosu (artemisinin): the price of success. *Science* 320(5874):330-334.

71. Organization WH (2006) *WHO Guidelines for the Safe Use of Wastewater, Excreta and Greywater* (World Health Organization, Geneva, Switzerland).

72. Barnes KI, *et al.* (2005) Effect of artemether-lumefantrine policy and improved vector control on malaria burden in KwaZulu–Natal, South Africa. *PLoS Med.* 2(11):e330.

73. Bhattarai A, *et al.* (2007) Impact of artemisinin-based combination therapy and insecticide-treated nets on malaria burden in Zanzibar. *PLoS Med.* 4(11):e309.

74. Carrara VI, *et al.* (2006) Deployment of early diagnosis and mefloquine-artesunate treatment of falciparum malaria in Thailand: the Tak Malaria Initiative. *PLoS Med.* 3(6):e183.

75. Jefford CW (2001) Why artemisinin and certain synthetic peroxides are potent antimalarials. Implications for the mode of action. *Curr Med Chem.* 8(15):1803-1826.

76. Pandey AV, Tekwani BL, Singh RL, & Chauhan VS (1999) Artemisinin, an endoperoxide antimalarial, disrupts the hemoglobin catabolism and heme detoxification systems in malarial parasite. *J Biol Chem.* 274(27):19383-19388.

77. Wilcox M, *et al.* (2004) Artemisia annua as a traditional herbal antimalarial. *Traditional herbal medicines for modern times*, (CRC Press, Boca Raton, FL), pp 43-60.

78. Jansen FH (2006) The herbal tea approach for artemisinin as a therapy for malaria? *Trans Royal Soc Trop Med Hyg.* 100(3):285-286.

79. Räth K, *et al.* (2004) Pharmacokinetic study of artemisinin after oral intake of a traditional preparation of Artemisia annua L.(annual wormwood). *Am J Trop Med Hyg.* 70(2):128-132.

80. Haynes RK (2006) From artemisinin to new artemisinin antimalarials: biosynthesis, extraction, old and new derivatives, stereochemistry and medicinal chemistry requirements. *Curr Top Med Chem.* 6(5):509-537.

81. Weina PJ (2008) Artemisinins from folklore to modern medicine-transforming an herbal extract to life-saving drugs. *Parassitologia.* 50(1/2):25.

82. Suberu JO, *et al.* (2013) Anti-plasmodial polyvalent interactions in Artemisia annua L. aqueous extract–possible synergistic and resistance mechanisms. *PLoS One.* 8(11):e80790.

83. Elford BC, Roberts MF, Phillipson JD, & Wilson RJ (1987) Potentiation of the antimalarial activity of qinghaosu by methoxylated flavones. *Trans. R. Soc. Trop. Med. Hyg.* 81(3):434-436.

84. Liu KC-SC, Yang S-L, Roberts M, Elford B, & Phillipson J (1992) Antimalarial activity of Artemisia annua flavonoids from whole plants and cell cultures. *Plant Cell Rep.* 11(12):637-640.

85. Nemeikaitė-Čėnienė A, Imbrasaitė A, Sergedienė E, & Čėnas N (2005) Quantitative structure–activity relationships in prooxidant cytotoxicity of polyphenols: role of potential of phenoxyl radical/phenol redox couple. *Arch Biochem Biophys.* 441(2):182-190.

86. Yordi EG, Pérez EM, Matos MJ, & Villares EU (2012) Antioxidant and pro-oxidant effects of polyphenolic compounds and structure-activity relationship evidence. *Nutrition, well-being and health*, (InTech).

87. Barrett B, *et al.* (2010) Echinacea for treating the common cold: a randomized trial. *Ann Intern Med.* 153(12):769-777.

88. Jawad M, Schoop R, Suter A, Klein P, & Eccles R (2012) Safety and Efficacy Profile of Echinacea purpurea to Prevent Common Cold Episodes: A Randomized, Double-Blind, Placebo-Controlled Trial. *Evid Based Complement Alternat Med.* 2012:841315.

89. Smith T, *et al.* (2015) Herbal dietary supplement sales in US increase 6.8% in 2014. *HerbalGram.* 107:52-59.

90. Rotblatt M (2000) Herbal Medicine: Expanded Commission E Monographs. *Ann Intern Med.* 133(6):487-487.

91. Proksch A & Wagner H (1987) Structural analysis of a 4-O-methyl-glucuronoarabinoxylan with immuno-stimulating activity from Echinacea purpurea. *Phytochemistry.* 26(7):1989-1993.

92. Pugh ND, *et al.* (2008) The majority of in vitro macrophage activation exhibited by extracts of some immune enhancing botanicals is due to bacterial lipoproteins and lipopolysaccharides. *Int Immunopharmacol.* 8(7):1023-1032.

93. Tamta H, *et al.* (2008) Variability in in vitro macrophage activation by commercially diverse bulk echinacea plant material is predominantly due to bacterial lipoproteins and lipopolysaccharides. *J Ag Food Chem.* 56(22):10552-10556.

94. Haron MH, *et al.* (2016) Activities and prevalence of Proteobacteria members colonizing Echinacea purpurea fully account for macrophage activation exhibited by extracts of this botanical. *Planta Med.* 82(14):1258.

95. Todd DA, *et al.* (2015) Ethanolic Echinacea purpurea extracts contain a mixture of cytokine-suppressive and cytokine-inducing compounds, including some that originate from endophytic bacteria. *PloS One.* 10(5):e0124276.

96. Cech NB, *et al.* (2010) Echinacea and its alkylamides: Effects on the influenza A-induced secretion of cytokines, chemokines, and PGE2 from RAW 264.7 macrophage-like cells. *Int Immunopharmacol.* 10(10):1268-1278.

97. Gertsch J, Schoop R, Kuenzle U, & Suter A (2004) Echinacea alkylamides modulate TNF-α gene expression via cannabinoid receptor CB2 and multiple signal transduction pathways. *FEBS Lett.* 577(3):563-569.

98. Raduner S, *et al.* (2006) Alkylamides from Echinacea are a new class of cannabinomimetics Cannabinoid type 2 receptor-dependent and-independent immunomodulatory effects. *J Biol Chem.* 281(20):14192-14206.

99. Stevenson L, *et al.* (2005) Modulation of macrophage immune responses by Echinacea. *Molecules.* 10(10):1279-1285.

100. Hemaiswarya S, Kruthiventi AK, & Doble M (2008) Synergism between natural products and antibiotics against infectious diseases. *Phytomedicine.* 15(8):639-652.

101. Gong X & Sucher NJ (1999) Stroke therapy in traditional Chinese medicine (TCM): prospects for drug discovery and development. *Trends Pharmacol Sci.* 20(5):191-196.

102. Brooks BD & Brooks AE (2014) Therapeutic strategies to combat antibiotic resistance. *Adv Drug Deliv Rev.* 78:14-27.

103. Hu C-MJ & Zhang L (2012) Nanoparticle-based combination therapy toward overcoming drug resistance in cancer. *Biochem Pharmacol.* 83(8):1104-1111.

104. Carmona F & Pereira AMS (2013) Herbal medicines: old and new concepts, truths and misunderstandings. *Rev Bras Farmacogn.* 23(2):379-385.

105. Koehn FE & Carter GT (2005) The evolving role of natural products in drug discovery. *Nat Rev Drug Discov.* 4(3):206.

106. Lila MA & Raskin I (2005) Health-related interactions of phytochemicals. *J Food Sci.* 70(1):R20-R27.

107. Imming P, Sinning C, & Meyer A (2006) Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov.* 5(10):821.

108. Montes P, Ruiz-Sanchez E, Rojas C, & Rojas P (2015) Ginkgo biloba Extract 761: A Review of Basic Studies and Potential Clinical Use in Psychiatric Disorders. *CNS Neurol Disord Drug Targets.* 14(1):132-149.

109. Spinella M (2002) The importance of pharmacological synergy in psychoactive herbal medicines. *Altern Med Rev.* 7(2):130-137.

110. Butterweck V, Liefländer-Wulf U, Winterhoff H, & Nahrstedt A (2003) Plasma levels of hypericin in presence of procyanidin B2 and hyperoside: a pharmacokinetic study in rats. *Planta Med.* 69(03):189-192.

111. Adams M, Mahringer A, Bauer R, Fricker G, & Efferth T (2007) In vitro cytotoxicity and P-glycoprotein modulating effects of geranylated furocoumarins from Tetradium daniellii. *Planta Med.* 73(14):1475-1478.

112. Adams M, *et al.* (2007) Cytotoxicity and p-glycoprotein modulating effects of quinolones and indoloquinazolines from the Chinese herb Evodia rutaecarpa. *Planta Med.* 73(15):1554-1557.

113. Liang X-L, *et al.* (2012) The absorption characterization effects and mechanism of Radix Angelicae dahuricae extracts on baicalin in Radix Scutellariae using in vivo and in vitro absorption models. *J Ethnopharmacol.* 139(1):52-57.

114. Wang S, Zhu F, & Marcone MF (2015) Staghorn sumac reduces 5-fluorouracil-induced toxicity in normal cells. *J Med Food.* 18(8):938-940.

115. McCune LM & Johns T (2002) Antioxidant activity in medicinal plants associated with the symptoms of diabetes mellitus used by the indigenous peoples of the North American boreal forest. *J Ethnopharmacol.* 82(2-3):197-205.

116. Müller R (2004) Crosstalk of oncogenic and prostanoid signaling pathways. *J Cancer Res Clin Oncol.* 130(8):429-444.

117. Kassouf W, *et al.* (2005) Uncoupling between epidermal growth factor receptor and downstream signals defines resistance to the antiproliferative effect of Gefitinib in bladder cancer cells. *Cancer Res.* 65(22):10524-10535.

118. Sergina NV, *et al.* (2007) Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. *Nature.* 445(7126):437-441.

119. Hu Q, Sun W, Wang C, & Gu Z (2016) Recent advances of cocktail chemotherapy by combination drug delivery systems. *Adv Drug Deliv Rev.* 98:19-34.

120. Sanglard D (2016) Emerging Threats in Antifungal-Resistant Fungal Pathogens. *Front Med.* 3:11-11.

121. Strasfeld L & Chou S (2010) Antiviral drug resistance: mechanisms and clinical implications. *Infect Dis Clin North Am.* 24(2):413-437.

122. O'Neill J (2016) Tackling Drug-Resistant Infections Globally: Final Report and Recommendations-The Review on Antimicrobial Resistance.

123. Jacoby GA & Sutton L (1985) beta-Lactamases and beta-lactam resistance in Escherichia coli. *Antimicrob Agents Chemother.* 28(5):703-705.

124. Catteau L, Olson J, Van Bambeke F, Leclercq J, & Nizet V (2017) Ursolic acid from shea butter tree (Vitellaria paradoxa) leaf extract synergizes with β-lactams against methicillin-resistant Staphylococcus aureus. *FASEB J.* 31(1_supplement):1000.1005-1000.1005.

125. Falagas ME & Bliziotis IA (2007) Pandrug-resistant Gram-negative bacteria: the dawn of the post-antibiotic era? *Int J Antimicrob Agents* 29(6):630-636.

126. Piddock LJ (2006) Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria. *Clin Microbiol Rev.* 19(2):382-402.

127. Ettefagh KA, Burns JT, Junio HA, Kaatz GW, & Cech NB (2011) Goldenseal (Hydrastis canadensis L.) extracts synergistically enhance the antibacterial activity of berberine via efflux pump inhibition. *Planta Med.* 77(8):835.

128. Leyte-Lugo M, *et al.* (2017) Secondary metabolites from the leaves of the medicinal plant goldenseal (Hydrastis canadensis). *Phytochem Lett.* 20:54-60.

129. Khan IA (2006) Issues related to botanicals. *Life Sci.* 78(18):2033-2038.

130. Bucar F, Wube A, & Schmid M (2013) Natural product isolation--how to get from biological material to pure compounds. *Nat Prod Rep.* 30(4):525-545.

131. Inui T, Wang Y, Pro SM, Franzblau SG, & Pauli GF (2012) Unbiased evaluation of bioactive secondary metabolites in complex matrices. *Fitoterapia.* 83(7):1218-1225.

132. Kellogg JJ, *et al.* (2016) Biochemometrics for natural products research: comparison of data analysis approaches and application to identification of bioactive compounds. *J Nat Prod.* 79(2):376-386.

133. Kurita KL, Glassey E, & Linington RG (2015) Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. *Proc Natl Acad Sci.* 112(39):11999-12004.

134. Nothias LF, *et al.* (2018) Bioactivity-Based Molecular Networking for the Discovery of Drug Leads in Natural Product Bioassay-Guided Fractionation. *J Nat Prod.* 81(4):758-767.

135. Oberlies NH & Kroll DJ (2004) Camptothecin and taxol: historic achievements in natural products research. *J Nat Prod.* 67(2):129-135.

136. El-Elimat T, *et al.* (2013) High-resolution MS, MS/MS, and UV database of fungal secondary metabolites as a dereplication protocol for bioactive natural products. *J Nat Prod.* 76(9):1709-1716.

137. Gaudencio SP & Pereira F (2015) Dereplication: racing to speed up the natural products discovery process. *Nat Prod Rep.* 32(6):779-810.

138. Wang M, *et al.* (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol.* 34(8):828-837.

139. Covington BC, McLean JA, & Bachmann BO (2017) Comparative mass spectrometry-based metabolomics strategies for the investigation of microbial secondary metabolites. *Nat Prod Rep.* 34(1):6-24.

140. Henke MT & Kelleher NL (2016) Modern mass spectrometry for synthetic biology and structure-based discovery of natural products. *Nat Prod Rep.* 33(8):942-950.

141. Kind T & Fiehn O (2017) Strategies for dereplication of natural compounds using high-resolution tandem mass spectrometry. *Phytochem Lett.* 21:313-319.

142. Williams RB, *et al.* (2015) Dereplication of natural products using minimal NMR data inputs. *Org Biomol Chem.* 13(39):9957-9962.

143. Yang JY, *et al.* (2013) Molecular networking as a dereplication strategy. *J Nat Prod.* 76(9):1686-1699.

144. Caesar LK, Kellogg JJ, Kvalheim OM, & Cech NB (2019) Opportunities and Limitations for Untargeted Mass Spectrometry metabolomics to Identify Biologically Active Constituents in Complex Natural Product Mixtures. *J. Nat. Prod.* :*in press.*

145. Caesar LK, Kellogg JJ, Kvalheim OM, Cech RA, & Cech NB (2018) Integration of Biochemometrics and Molecular Networking to Identify Antimicrobials in Angelica keiskei. *Planta Med.*

146. Chan KM, *et al.* (2017) Screening and analysis of potential anti-tumor components from the stipe of Ganoderma sinense using high-performance liquid chromatography/time-of-flight mass spectrometry with multivariate statistical tool. *J Chromatogr A.* 1487:162-167.

147. Li P, *et al.* (2016) Comparative UPLC-QTOF-MS-based metabolomics and bioactivities analyses of Garcinia oblongifolia. *J Chromatogr B.* 1011:179-195.

148. Abdi H (2004) Partial Least Squares Regression. *The SAGE Encyclopedia of Social Science Research Methods*, (SAGE Publications, Inc., Thousand Oaks, CA), pp 792-795.

149. Rajalahti T, *et al.* (2009) Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Anal Chem.* 81(7):2581-2590.

150. Rajalahti T & Kvalheim OM (2011) Multivariate data analysis in pharmaceutics: a tutorial review. *Int J Pharm.* 417(1-2):280-290.

151. Chen S, *et al.* (2016) Drug target identification using network analysis: Taking active components in Sini decoction as an example. *Sci Rep.* 6:24245-24245.

152. Brown AR, *et al.* (2015) A Mass Spectrometry-Based Assay for Improved Quantitative Measurements of Efflux Pump Inhibition. *PLoS One* 10(5):e0124814.

153. Blair JMA & Piddock LJV (2016) How to Measure Export via Bacterial Multidrug Resistance Efflux Pumps. *mBio* 7(4):e00840-00816.

154. Bohnert JA, Karamian B, & Nikaido H (2010) Optimized Nile Red efflux assay of AcrAB-TolC multidrug efflux system shows competition between substrates. *Antimicrob Agents Chemother.* 54(9):3770-3775.

155. Baell JB (2016) Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). *J Nat Prod.* 79(3):616-628.

156. Baell JB & Holloway GA (2010) New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J Med Chem.* 53(7):2719-2740.

157. Davis TD, Gerry CJ, & Tan DS (2014) General Platform for Systematic Quantitative Evaluation of Small-Molecule Permeability in Bacteria. *ACS Chem Biol.* 9(11):2535-2544.

158. Richter MF, *et al.* (2017) Predictive compound accumulation rules yield a broad-spectrum antibiotic. *Nature.* 545(7654):299-304.

159. van Breemen RB, *et al.* (1997) Pulsed Ultrafiltration Mass Spectrometry: A New Method for Screening Combinatorial Libraries. *Anal Chem.* 69(11):2159-2164.

160. Rush MD, Walker EM, Prehna G, Burton T, & van Breemen RB (2017) Development of a Magnetic Microbead Affinity Selection Screen (MagMASS) Using Mass Spectrometry for Ligands to the Retinoid X Receptor-alpha. *J Am Soc Mass Spectrom.* 28(3):479-485.

161. Panossian A, Seo EJ, Wikman G, & Efferth T (2015) Synergy assessment of fixed combinations of Herba Andrographidis and Radix Eleutherococci extracts by transcriptome-wide microarray profiling. *Phytomedicine.* 22(11):981-992.

162. Xiong Y, *et al.* (2019) Unveiling Active Constituents and Potential Targets Related to the Hematinic Effect of Steamed Panax notoginseng Using Network Pharmacology Coupled With Multivariate Data Analyses. *Front Pharmacol.* 9(1514).

163. Potts MB, *et al.* (2013) Using functional signature ontology (FUSION) to identify mechanisms of action for natural products. *Sci Signal.* 6(297):ra90.

164. Das B, *et al.* (2018) A Functional Signature Ontology (FUSION) screen detects an AMPK inhibitor with selective toxicity toward human colon tumor cells. *Sci Rep.* 8(1):3770-3770.

165. Subramanian A, *et al.* (2017) A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell.* 171(6):1437-1452.e1417.

166. Sarker S & Nahar L (2004) Natural medicine: the genus Angelica. *Curr Med Chem.* 11(11):1479-1500.

167. Akihisa T, *et al.* (2006) Chalcones and Other Compounds from the Exudates of Angelica k eiskei and Their Cancer Chemopreventive Effects. *J Nat Prod.* 69(1):38-42.

168. Kim DW, *et al.* (2014) Quantitative analysis of phenolic metabolites from different parts of Angelica keiskei by HPLC–ESI MS/MS and their xanthine oxidase inhibition. *Food Chem.* 153:20-27.

169. Winkel-Shirley B (2001) Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* 126(2):485-493.

170. Battenberg OA, Yang Y, Verhelst SH, & Sieber SA (2013) Target profiling of 4-hydroxyderricin in S. aureus reveals seryl-tRNA synthetase binding and inhibition by covalent modification. *Mol Biosyst.* 9(3):343-351.

171. Mahapatra DK, Asati V, & Bharti SK (2015) Chalcones and their therapeutic targets for the management of diabetes: structural and pharmacological perspectives. *Eur J Med Chem.* 92:839-865.

172. McCall S, Don Johnson R, & Col U (2013) Investigation of the anxiolytic effects of xanthohumol, a component of humulus lupulus (Hops), in the male Sprague-Dawley rat. *AANA J.* 81(3):193.

173. Yadav VR, Prasad S, Sung B, & Aggarwal BB (2011) The role of chalcones in suppression of NF-κB-mediated inflammation and cancer. *Int Immunopharmacol.* 11(3):295-309.

174. Zhang H, *et al.* (2012) Design, synthesis and biological evaluation of novel chalcone derivatives as antitubulin agents. *Bioorg Med Chem.* 20(10):3212-3218.

175. Zhou B & Xing C (2015) Diverse molecular targets for chalcones with varied bioactivities. *Med Chem.* 5(8):388.

176. Zhang T, Yamashita Y, Yasuda M, Yamamoto N, & Ashida H (2015) Ashitaba (Angelica keiskei) extract prevents adiposity in high-fat diet-fed C57BL/6 mice. *Food Funct.* 6(1):134-144.

177. Kozawa M, Morita N, Baba K, & Hata K (1977) The structure of xanthoangelol, a new chalcone from the roots of Angelica keiskei Koidzumi (Umbelliferae). *Chem Pharm Bull.* 25(3):515-516.

178. Baba K, Nakata K, Taniguchi M, Kido T, & Kozawa M (1990) Chalcones from Angelica keiskei. *Phytochemistry* 29(12):3907-3910.

179. Nakata K, Taniguchi M, & Baba K (1999) Three chalcones from Angelica keiskei. *Natural Med.* 53(6):329-332.

180. Li J-l, *et al.* (2015) PTP1B inhibitors from stems of Angelica keiskei (Ashitaba). *Bioorg Med Chem Lett.* 25(10):2028-2032.

181. Park J-Y, *et al.* (2011) Characteristic of alkylated chalcones from Angelica keiskei on influenza virus neuraminidase inhibition. *Bioorg Med Chem Lett.* 21(18):5602-5604.

182. Ohnogi H, *et al.* (2012) Six new chalcones from Angelica keiskei inducing adiponectin production in 3T3-L1 adipocytes. *Biosci Biotechnol Biochem.* 76(5):961-966.

183. Aoki N, Muko M, Ohta E, & Ohta S (2008) C-geranylated chalcones from the stems of Angelica keiskei with superoxide-scavenging activity. *J Nat Prod.* 71(7):1308-1310.

184. Akihisa T, *et al.* (2003) Chalcones, coumarins, and flavanones from the exudate of Angelica keiskei and their chemopreventive effects. *Cancer Lett.* 201(2):133-137.

185. Luo L, Wang R, Wang X, Ma Z, & Li N (2012) Compounds from Angelica keiskei with NQO1 induction, DPPH scavenging and α-glucosidase inhibitory activities. *Food Chem.* 131(3):992-998.

186. Aoki N & Ohta S (2010) Ashitabaol A, a new antioxidative sesquiterpenoid from seeds of Angelica keiskei. *Tetrahedron Lett.* 51(26):3449-3450.

187. Bourgaud F, *et al.* (2006) Biosynthesis of coumarins in plants: a major pathway still to be unravelled for cytochrome P450 enzymes. *Phytochem Rev.* 5(2-3):293-308.

188. Fylaktakidou KC, Hadjipavlou-Litina DJ, Litinas KE, & Nicolaides DN (2004) Natural and synthetic coumarin derivatives with anti-inflammatory/antioxidant activities. *Curr Pharm Des.* 10(30):3813-3833.

189. Musa MA, Cooperwood JS, & Khan MOF (2008) A review of coumarin derivatives in pharmacotherapy of breast cancer. *Curr Med Chem.* 15(26):2664-2679.

190. Okuyama T, *et al.* (1991) Anti-tumor-promotion by principles obtained from Angelica keiskei. *Planta Med.* 57(03):242-246.

191. Ogawa H, Nakamura R, & Baba K (2005) Beneficial effect of laserpitin, a coumarin compound from Angelica keiskei, on lipid metabolism in stroke-prone spontaneously hypertensive rats. *Clin Exp Pharmacol Physiol.* 32(12):1104-1109.

192. Wong E (1968) The role of chalcones and flavanones in flavonoid biosynthesis. *Phytochemistry.* 7(10):1751-1758.

193. Khan MK & Dangles O (2014) A comprehensive review on flavanones, the major citrus polyphenols. *J Food Comp Anal.* 33(1):85-104.

194. Hashimoto K, Kawamata S, Usui N, Tanaka A, & Uda Y (2002) In vitro induction of the anticarcinogenic marker enzyme, quinone reductase, in human hepatoma cells by food extracts. *Cancer Lett.* 180(1):1-5.

195. Ogawa H, Nakashima S, & Baba K (2003) Effects of dietary Angelica keiskei on lipid metabolism in stroke-prone spontaneously hypertensive rats. *Clin Exp Pharmacol Physiol.* 30(4):284-288.

196. Kim E, Choi J, & Yeo I (2012) The effects of Angelica keiskei Koidz on the expression of antioxidant enzymes related to lipid profiles in rats fed a high fat diet. *Nutr Res Pract.* 6(1):9-15.

197. Nagata J, Morino T, & Saito M (2007) Effects of dietary Angelica keiskei on serum and liver lipid profiles, and body fat accumulations in rats. *J Nutr Sci Vitaminol.* 53(2):133-137.

198. Ohnogi H, *et al.* (2012) Angelica keiskei extract improves insulin resistance and hypertriglyceridemia in rats fed a high-fructose drink. *Biosci Biotechnol Biochem.* 76(5):928-932.

199. Enoki T, *et al.* (2007) Antidiabetic activities of chalcones isolated from a Japanese herb, Angelica keiskei. *J Ag Food Chem.* 55(15):6013-6017.

200. Lee HJ, *et al.* (2010) Anti-Inflammatory Activity of Angelica keiskei Through Suppression of Mitogen-Activated Protein Kinases and Nuclear Factor-κ B Activation Pathways. *J Med Food.* 13(3):691-699.

201. Ohkura N, *et al.* (2011) Xanthoangelols isolated from Angelica keiskei inhibit inflammatory-induced plasminogen activator inhibitor 1 (PAI-1) production. *Biofactors.* 37(6):455-461.

202. Shimizu E, *et al.* (1999) Effects of angiotensin I-converting enzyme inhibitor from Ashitaba (Angelica keiskei) on blood pressure of spontaneously hypertensive rats. *J Nutr Sci Vitaminol.* 45(3):375-383.

203. Akihisa T, *et al.* (2012) Cytotoxic Activities and Anti-Tumor-Promoting Effects of Microbial Transformation Products of Prenylated Chalcones from Angelica keiskei. *Chem Biodivers.* 9(2):318-330.

204. Inamori Y, *et al.* (1991) Antibacterial activity of two chalcones, xanthoangelol and 4-hydroxyderricin, isolated from the root of Angelica keiskei KOIDZUMI. *Chem Pharm Bull.* 39(6):1604-1605.

205. Kawabata K, *et al.* (2011) Prenylated chalcones 4-hydroxyderricin and xanthoangelol stimulate glucose uptake in skeletal muscle cells by inducing GLUT4 translocation. *Mol Nut Food Res.* 55(3):467-475.

206. Son DJ, Park YO, Yu C, Lee SE, & Park YH (2014) Bioassay-guided isolation and identification of anti-platelet-active compounds from the root of Ashitaba (Angelica keiskei Koidz.). *Nat Prod Res.* 28(24):2312-2316.

207. Sugii M, *et al.* (2005) Xanthoangelol D isolated from the roots of Angelica keiskei inhibits endothelin-1 production through the suppression of nuclear factor-κB. *Biol Pharm Bull.* 28(4):607-610.

208. Yasuda M, *et al.* (2014) Inhibitory effects of 4-hydroxyderricin and xanthoangelol on lipopolysaccharide-induced inflammatory responses in RAW264 macrophages. *J Ag Food Chem.* 62(2):462-467.

209. Zhang T, Sawada K, Yamamoto N, & Ashida H (2013) 4-Hydroxyderricin and xanthoangelol from Ashitaba (Angelica keiskei) suppress differentiation of preadiopocytes to adipocytes via AMPK and MAPK pathways. *Mol Nut Food Res.* 57(10):1729-1740.

210. Nakata K & Baba K (2001) Histamine release inhibition activity of Angelica keiskei. *Natural Med.* 55(1):32-34.

211. Ohnogi H, Hayami S, Kudo Y, & Enoki T (2012) Efficacy and safety of ashitaba (Angelica keiskei) on the patients and candidates with metabolic syndrome: a pilot study. *JCAM.* 9(1):49-55.

212. Bolca S, *et al.* (2010) Disposition of hop prenylflavonoids in human breast tissue. *Mol Nut Food Res.* 54(S2):S284-S294.

213. Hanske L, Loh G, Sczesny S, Blaut M, & Braune A (2010) Recovery and metabolism of xanthohumol in germ-free and human microbiota-associated rats. *Mol Nut Food Res.* 54(10):1405-1413.

214. Wesołowska O, Gąsiorowska J, Petrus J, Czarnik-Matusewicz B, & Michalak K (2014) Interaction of prenylated chalcones and flavanones from common hop with phosphatidylcholine model membranes. *Biochim Biophys Acta.* 1838(1):173-184.

215. Maronpot RR (2015) Toxicological assessment of ashitaba chalcone. *Food Chem Toxicol.* 77:111-119.

216. Son H-U, *et al.* (2012) Comparison of the toxicity of aqueous and ethanol fractions of Angelica keiskei leaf using the eye irritancy test. *Exp Ther Med.* 4(5):820-824.

217. Vignes S & Bellanger J (2008) Primary intestinal lymphangiectasia (Waldmann's disease). *Orphanet J Rare Dis.* 3(1):5.

218. Messer A, Raquet N, Lohr C, & Schrenk D (2012) Major furocoumarins in grapefruit juice II: phototoxicity, photogenotoxicity, and inhibitory potency vs. cytochrome P450 3A4 activity. *Food Chem Toxicol.* 50(3-4):756-760.

219. Eisenbrand G (2007) Toxicological assessment of furocoumarins in foodstuffs. *Mol Nut Food Res.* 51(3):367-373.

220. Kinghorn AD (1998) Cancer chemopreventive agents discovered by activity-guided fractionation. *Curr Org Chem.* 2(6):597-612.

221. Sharma SB & Gupta R (2015) Drug development from natural resource: a systematic approach. *Mini Rev Med Chem.* 15(1):52-57.

222. Roemer T, *et al.* (2011) Confronting the challenges of natural product-based antifungal discovery. *Chem Biol.* 18(2):148-164.

223. Newman DJ & Cragg GM (2016) Natural Products as Sources of New Drugs from 1981 to 2014. *J Nat Prod.* 79(3):629-661.

224. Cowan MM (1999) Plant products as antimicrobial agents. *Clin Microbiol Rev.* 12(4):564-582.

225. Gurjar MS, Ali S, Akhtar M, & Singh KS (2012) Efficacy of plant extracts in plant disease management. *Ag Sci.* 3(3):425.

226. Kingston DG (2011) Modern natural products drug discovery and its relevance to biodiversity conservation. *J Nat Prod.* 74(3):496-511.

227. Savoia D (2012) Plant-derived antimicrobial compounds: alternatives to antibiotics. *Fut Microbiol.* 7(8):979-990.

228. Caesar LK & Cech NB (2016) A review of the medicinal uses and pharmacology of ashitaba. *Planta Med.* 82(14):1236-1245.

229. Sugamoto K, Matsusita Y-i, Matsui K, Kurogi C, & Matsui T (2011) Synthesis and antibacterial activity of chalcones bearing prenyl or geranyl groups from Angelica keiskei. *Tetrahedron.* 67(29):5346-5359.

230. Sidebottom AM, Johnson AR, Karty JA, Trader DJ, & Carlson EE (2013) Integrated metabolomics approach facilitates discovery of an unpredicted natural product suite from Streptomyces coelicolor M145. *ACS Chem Biol.* 8(9):2009-2016.

231. Wu C, Du C, Gubbens J, Choi YH, & van Wezel GP (2015) Metabolomics-driven discovery of a prenylated isatin antibiotic produced by streptomyces species MBT28. *J Nat Prod.* 78(10):2355-2363.

232. Cox DG, Oh J, Keasling A, Colson KL, & Hamann MT (2014) The utility of metabolomics in natural product and biomarker characterization. *Biochim Biophys Acta.* 1840(12):3460-3474.

233. Rajalahti T, *et al.* (2009) Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemom Int Lab Syst.* 95(1):35-48.

234. Boles BR, Thoendel M, Roth AJ, & Horswill AR (2010) Identification of Genes Involved in Polysaccharide-Independent Staphylococcus aureus Biofilm Formation. *PLoS One* 5(4):e10146.

235. Kim JH, Son YK, Kim GH, & Hwang KH (2013) Xanthoangelol and 4-hydroxyderricin are the major active principles of the inhibitory activities against monoamine oxidases on Angelica keiskei K. *Biomol Ther.* 21(3):234.

236. CLSI (2015) Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria that Grow Aerobically--Tenth Edition:Approved Standard M7-A10. *National Committee for Clinical Laboratory Standards*, Wayne, PA).

237. Nowakowska Z (2007) A review of anti-infective and anti-inflammatory chalcones. *Eur J Med Chem.* 42(2):125-137.

238. Kaatz GW & Seo SM (1995) Inducible NorA-mediated multidrug resistance in Staphylococcus aureus. *Antimicrob Agents Chemother.* 39(12):2650-2655.

239. Pluskal T, Castillo S, Villar-Briones A, & Orešič M (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics.* 11(1):395.

240. Kvalheim OM, *et al.* (2011) Chromatographic profiling and multivariate analysis for screening and quantifying the contributions from individual components to the bioactive signature in natural products. *Chemom Int Lab Syst.* 107(1):98-105.

241. Kvalheim OM, Brakstad F, & Liang Y (1994) Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Anal Chem.* 66(1):43-51.

242. Frank AM, Savitski MM, Nielsen ML, Zubarev RA, & Pevzner PA (2007) De novo peptide sequencing and identification with precision mass spectrometry. *J Proteom Res.* 6(1):114-123.

243. Shannon P, *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13(11):2498-2504.

244. Boccard J, Veuthey JL, & Rudaz S (2010) Knowledge discovery in metabolomics: an overview of MS data handling. *J Sep Sci.* 33(3):290-304.

245. Ceglarek U, *et al.* (2009) Challenges and developments in tandem mass spectrometry based clinical metabolomics. *Mol Cell Endocrinol.* 301(1-2):266-271.

246. Klupczyńska A, Dereziński P, & Kokot ZJ (2015) Metabolomics in Medical Sciences--Trends, Challenges, and Perspectives. *Acta Pol Pharm.* 72(4):629-641.

247. Li X, *et al.* (2017) Metabolic characterization and pathway analysis of berberine protects against prostate cancer. *Oncotarget.* 8(39):65022-65041.

248. Matsuda F (2016) Technical Challenges in Mass Spectrometry-Based Metabolomics. *Mass Spectrom (Tokyo).* 5(2):S0052.

249. Monteiro M, Carvalho M, Bastos M, & Guedes de Pinho P (2013) Metabolomics analysis for biomarker discovery: advances and challenges. *Curr Med Chem.* 20(2):257-271.

250. Wishart DS (2008) Metabolomics: applications to food science and nutrition research. *Trends Food Sci Technol.* 19(9):482-493.

251. Ackermann BL, Hale JE, & Duffin KL (2006) The role of mass spectrometry in biomarker discovery and measurement. *Curr Drug Metab.* 7(5):525-539.

252. Sreekumar A, *et al.* (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature.* 457(7231):910-914.

253. van der Greef J, Hankemeier T, & McBurney RN (2006) Metabolomics-based systems biology and personalized medicine: moving towards n = 1 clinical trials? *Pharmacogenomics.* 7(7):1087-1094.

254. Chau F-T, *et al.* (2009) Recipe for uncovering the bioactive components in herbal medicine. *Anal Chem.* 81(17):7217-7225.

255. Mikami T, Aoki M, & Kimura T (2012) The application of mass spectrometry to proteomics and metabolomics in biomarker discovery and drug development. *Curr Mol Pharm.* 5(2):301-316.

256. Marshall DD & Powers R (2017) Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics. *Prog Nucl Magn Reson Spectrosc.* 100:1-16.

257. Rochfort S (2005) Metabolomics reviewed: a new "omics" platform technology for systems biology and implications for natural products research. *J Nat Prod.* 68(12):1813-1820.

258. de Jong FA & Beecher C (2012) Addressing the current bottlenecks of metabolomics: Isotopic Ratio Outlier Analysis™, an isotopic-labeling technique for accurate biochemical profiling. *Bioanalysis* 4(18):2303-2314.

259. McMillan A, Renaud JB, Gloor GB, Reid G, & Sumarah MW (2016) Post-acquisition filtering of salt cluster artefacts for LC-MS based human metabolomic studies. *J Cheminformatics.* 8(1):44.

260. Majors RE (2006) Developments in HPLC column packing design. *LCGC North America.* 24(4):8-15.

261. Wyndham K, *et al.* (2004) *A review of waters hybrid particle technology, part 2* (Waters Corporation, Milford, Milford).

262. Berg M, *et al.* (2013) LC-MS metabolomics from study design to data-analysis–using a versatile pathogen as a test case. *Comp Struct Biol J.* 4(5):e201301002.

263. Aretz I & Meierhofer D (2016) Advantages and pitfalls of mass spectrometry based metabolome profiling in systems biology. *Int J Mol Sci.* 17(5):632.

264. Moran NE, *et al.* (2013) Biosynthesis of highly enriched 13C-lycopene for human metabolic studies using repeated batch tomato cell culturing with 13C-glucose. *Food Chem.* 139(1-4):631-639.

265. Cano PM, *et al.* (2013) New untargeted metabolic profiling combining mass spectrometry and isotopic labeling: application on Aspergillus fumigatus grown on wheat. *Anal Chem.* 85(17):8412-8420.

266. Eisen MB, Spellman PT, Brown PO, & Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci.* 95(25):14863-14868.

267. Beckonert O, *et al.* (2003) NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. *Anal Chim Acta.* 490(1-2):3-15.

268. Tikunov Y, *et al.* (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol.* 139(3):1125-1137.

269. Kvalheim OM & Karstang TV (1989) Interpretation of latent-variable regression models. *Chemom Int Lab Syst.* 7(1-2):39-51.

270. Kaufman L & Rousseeuw PJ (2009) *Finding groups in data: an introduction to cluster analysis* (John Wiley & Sons).

271. Keller BO, Sui J, Young AB, & Whittal RM (2008) Interferences and contaminants encountered in modern mass spectrometry. *Anal Chim Acta.* 627(1):71-81.

272. Korman A, Oh A, Raskind A, & Banks D (2012) Statistical methods in metabolomics. *Methods Mol Biol.* 856:381-413.

273. Kuehnbaum NL & Britz-McKibbin P (2013) New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era. *Chem Rev.* 113(4):2437-2468.

274. Altenburger R, *et al.* (2015) Future water quality monitoring — Adapting tools to deal with mixtures of pollutants in water resource management. *Sci Total Environ.* 512-513:540-551.

275. Heinrich M (2008) *Ethnopharmacy and natural product research—Multidisciplinary opportunities for research in the metabolomic age*.

276. Caesar LK, Kvalheim OM, & Cech NB (2018) Hierarchical Cluster Analysis of Technical Replicates to Identify Interferents in Untargeted Mass Spectrometry Metabolomics. *Anal Chim Acta.* 1021.

277. Yuliana ND, Khatib A, Choi YH, & Verpoorte R (2011) Metabolomics for bioactivity assessment of natural products. *Phytother Res.* 25(2):157-169.

278. Okada T, Afendi FM, Katoh A, Hirai A, & Kanaya S (2013) Multivariate analysis of analytical chemistry data and utility of the KNApSAcK family database to understand metabolic diversity in medicinal plants. *Biotechnol Med Plants.*, (Springer), pp 413-438.

279. Wiklund S, *et al.* (2008) Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Anal Chem.* 80(1):115-122.

280. Yun X, Dong S, Hu Q, Dai Y, & Xia Y (2018) (1)H NMR-based metabolomics approach to investigate the urine samples of collagen-induced arthritis rats and the intervention of tetrandrine. *J Pharm Biomed Anal.* 154:302-311.

281. Hrbek V, Rektorisova M, Chmelarova H, Ovesna J, & Hajslova J (2018) Authenticity assessment of garlic using a metabolomic approach based on high resolution mass spectrometry. *J Food Comp Anal.* 67:19-28.

282. Kulakowski DM, Wu SB, Balick MJ, & Kennelly EJ (2014) Merging bioactivity with liquid chromatography-mass spectrometry-based chemometrics to identify minor immunomodulatory compounds from a Micronesian adaptogen, Phaleria nisidai. *J Chromatogr A.* 1364:74-82.

283. Shang N, *et al.* (2015) Novel Approach to Identify Potential Bioactive Plant Metabolites: Pharmacological and Metabolomics Analyses of Ethanol and Hot Water Extracts of Several Canadian Medicinal Plants of the Cree of Eeyou Istchee. *PLoS One.* 10(8):e0135721.

284. Syu WJ, Shen CC, Lu JJ, Lee GH, & Sun CM (2004) Antimicrobial and cytotoxic activities of neolignans from Magnolia officinalis. *Chem Biodivers.* 1(3):530-537.

285. Lee DS, Lee SH, Noh JG, & Hong SD (1999) Antibacterial activities of cryptotanshinone and dihydrotanshinone I from a medicinal herb, Salvia miltiorrhiza Bunge. *Biosci Biotechnol Biochem.* 63(12):2236-2239.

286. Phuong NTM, *et al.* (2017) Antibiofilm activity of α-mangostin extracted from Garcinia mangostana L. against Staphylococcus aureus. *Asian Pac J Trop Med.* 10(12):1154-1160.

287. Brereton RG (2014) A short history of chemometrics: a personal view. *J Chemom.* 28(10):749-760.

288. Westerhuis JA, *et al.* (2008) Assessment of PLSDA cross validation. *Metabolomics.* 4(1):81-89.

289. Kemsley EK, *et al.* (2007) Multivariate techniques and their application in nutrition: a metabolomics case study. *Br J Nut.* 98(1):1-14.

290. Shah D & Madden L (2004) Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology.* 94(1):33-43.

291. Rietjens M (1995) Reduction of error propagation due to normalization: Effect of error propagation and closure on spurious correlations. *Anal Chim Acta.* 316(2):205-215.

292. Arneberg R, *et al.* (2007) Pretreatment of mass spectral profiles: application to proteomic data. *Anal Chem.* 79(18):7014-7026.

293. Kathiravan G, Sureban SM, Sree HN, Bhuvaneshwari V, & Kramony E (2012) Isolation of anticancer drug TAXOL from Pestalotiopsis breviseta with apoptosis and B-Cell lymphoma protein docking studies. *J Basic Clin Pharm.* 4(1):14-19.

294. da Costa JP, Santos PSM, Vitorino R, Rocha-Santos T, & Duarte AC (2017) How low can you go? A current perspective on low-abundance proteomics. *TrAC Trends Anal Chem.* 93:171-182.

295. Hewitt SM, Dear J, & Star RA (2004) Discovery of protein biomarkers for renal diseases. *J Am Soc Nephrol.* 15(7):1677-1689.

296. Baggerly KA, *et al.* (2003) A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* 3(9):1667-1672.

297. Eliopolous GM MR (1996) Antibiotics in Laboratory Medicine. ed V I (Williams and Wilkins, Baltimore MD), 3 Ed, pp 330-396.

298. Tukey JW (1962) The Future of Data Analysis. *Ann Math Statist.* 33(1):1-67.

299. Wang L, Yuana K, Yu WW, & Wang J (2010) Evaluation and discrimination of cortex Magnoliae officinalis produced in Zhejiang Province (Wen-Hou-Po) by UPLC-DAD-TOF-MS fingerprint. *Nat Prod Comm.* 5(10):1631-1638.

300. Yahara S, Nishiyori T, Kohda A, Nohara T, & Nishioka I (1991) Isolation and Characterization of Phenolic Compounds from Magnoliae Cortex Produced in China. *Chem Pharm Bull.* 39(8):2024-2036.

301. Bell A (2005) Antimalarial drug synergism and antagonism: mechanistic and clinical significance. *FEMS Microbiol Lett.* 253(2):171-184.

302. Johnson MD, MacDougall C, Ostrosky-Zeichner L, Perfect JR, & Rex JH (2004) Combination antifungal therapy. *Antimicrob Agents Chemother.* 48(3):693-715.

303. Lambert R & Lambert R (2003) A model for the efficacy of combined inhibitors. *J Appl Microbiol.* 95(4):734-743.

304. Institute HE (1996) Background for the Complex Mixtures Program. in *Theoretical Approaches to Analyzing Complex Mixtures* (Health Effects Institute, Cambridge, MA), p 76.

305. Abreu NA & Taga ME (2016) Decoding molecular interactions in microbial communities. *FEMS Microbiol Rev* 40(5):648-663.

306. Laxminarayan R, *et al.* (2013) Antibiotic resistance—the need for global solutions. *Lancet Infect Dis.* 13(12):1057-1098.

307. Wagner H (2006) Multitarget therapy--the future of treatment for more than just functional dyspepsia. *Phytomedicine.* 13 Suppl 5:122-129.

308. Hartmann RW, Mark M, & Soldati F (1996) Inhibition of 5 alpha-reductase and aromatase by PHL-00801 (Prostatonin(R)), a combination of PY102 (Pygeum africanum) and UR102 (Urtica dioica) extracts. *Phytomedicine.* 3(2):121-128.

309. Butterweck V, Jurgenliemk G, Nahrstedt A, & Winterhoff H (2000) Flavonoids from Hypericum perforatum show antidepressant activity in the forced swimming test. *Planta Med.* 66(1):3-6.

310. Houghton PJ (2000) Use of small scale bioassays in the discovery of novel drugs from natural sources. *Phytother Res.* 14(6):419-423.

311. Wall ME & Wani MC (1996) Camptothecin and taxol: from discovery to clinic. *J Ethnopharmacol.* 51(1-3):239-254.

312. Qiu F, *et al.* (2013) Quantitative purity–activity relationships of natural products: the case of anti-tuberculosis active triterpenes from Oplopanax horridus. *J Nat Prod.* 76(3):413-419.

313. Cha J-D, *et al.* (2013) Synergistic effect between cryptotanshinone and antibiotics in oral pathogenic bacteria. *Adv Biosci Biotechnol.* 4(02):283.

314. Cha J-D, Lee J-H, Choi KM, Choi S-M, & Park JH (2014) Synergistic effect between cryptotanshinone and antibiotics against clinic methicillin and vancomycin-resistant Staphylococcus aureus. *Evid. Based. Complement. Alternat. Med.* 2014.

315. Jamshidi-Aidji M & Morlock GE (2016) From bioprofiling and characterization to bioquantification of natural antibiotics by direct bioautography linked to high-resolution mass spectrometry: Exemplarily shown for Salvia miltiorrhiza root. *Anal Chem.* 88(22):10979-10986.

316. Lee J-W, Ji Y-J, Lee S-O, & Lee I-S (2007) Effect of Saliva miltiorrhiza bunge on antimicrobial activity and resistant gene regulation against methicillin-resistant Staphylococcus aureus (MRSA). *J Microbiol.* 45(4):350-357.

317. Wang B-Q (2010) Salvia miltiorrhiza: Chemical and pharmacological review of a medicinal plant. *J Med Plant Res.* 4(25):2813-2820.

318. Xu S & Liu P (2013) Tanshinone II-A: new perspectives for old remedies. *Expert Opin Ther Pat.* 23(2):149-153.

319. Cottarel G & Wierzbowski J (2007) Combination drugs, an emerging option for antibacterial therapy. *Trends Biotechnol.* 25(12):547-555.

320. Goñi P, *et al.* (2009) Antimicrobial activity in the vapour phase of a combination of cinnamon and clove essential oils. *Food Chem.* 116(4):982-989.

321. Testing ECfAS (2000) EUCAST Definitive Document E.Def 1.2, May 2000: Terminology relating to methods for the determination of susceptibility of bacteria to antimicrobial agents. *Clin Microbiol Infect* 6(9):503-508.

322. Li A, She X, Zhang J, Wu T, & Pan X (2003) Synthesis of C-7 oxidized abietane diterpenes from racemic ferruginyl methyl ether. *Tetrahedron.* 59(30):5737-5741.

323. Chang HM, *et al.* (1990) Structure elucidation and total synthesis of new tanshinones isolated from Salvia miltiorrhiza Bunge (Danshen). *J Org Chem.* 55(11):3537-3543.

324. Snyder SA (2014) Emerging chemical contaminants: looking for greater harmony. *J Am Water Works Assoc.* 106(8):38-52.

325. Delfosse V, *et al.* (2015) Synergistic activation of human pregnane X receptor by binary cocktails of pharmaceutical and environmental compounds. *Nat Comm.* 6:8089.

326. Kortenkamp A (2014) Low dose mixture effects of endocrine disrupters and their implications for regulatory thresholds in chemical risk assessment. *Curr Opin Pharmacol.* 19:105-111.

327. Ara I, Siddiqui BS, Faizi S, & Siddiqui S (1990) Tricyclic diterpenes from the stem bark of Azadirachta indica. *Planta Med.* 56(01):84-86.

328. Chao K-P, Hua K-F, Hsu H-Y, Su Y-C, & Chang S-T (2005) Anti-inflammatory activity of sugiol, a diterpene isolated from Calocedrus formosana bark. *Planta Med.* 71(04):300-305.

329. Dávila M, Stenier O, & Hinojosa N (2014) SECONDARY METABOLITES FROM PODOCARPUSPARLATOREI PILGER. *Revista Boliviana de Química* 31(1):22-28.

330. Gao J & Han G (1997) Cytotoxic abietane diterpenoids from Caryopteris incana. *Phytochemistry* 44(4):759-761.

331. Rodríguez B (2003) 1H and 13C NMR spectral assignments of some natural abietane diterpenoids. *Magn Reson Chem.* 41(9):741-746.

332. Wang S-Y, Wu J-H, Shyur L-F, Kuo Y-H, & Chang S-T (2002) Antioxidant activity of abietane-type diterpenes from heartwood of Taiwania cryptomerioides Hayata. *Holzforschung.* 56(5):487-492.

333. Kang HS, Chung HY, Jung JH, Kang SS, & Choi JS (1997) Antioxidant effect ofSalvia miltiorrhiza. *Arch Pharmacal Res.* 20(5):496.

334. Sairafianpour M, *et al.* (2001) Leishmanicidal, Antiplasmodial, and Cytotoxic Activity of Novel Diterpenoid 1,2Quinones from Perovskia abrotanoides : New Source of Tanshinones. *J Nat Prod.* 64:1398-1403.

APPENDIX A

SUPPLEMENTARY PROTOCOLS

Protocol S1: Detailed Sample Preparation Procedure to Produce Samples for Hierarchical Cluster Analysis

Protocol S2: Plant Extraction and Simplification of *Angelica keiskei* fraction

Protocol S3: Chromatographic Separation and Isolation of *Salvia miltiorrhiza*

Protocol S1. Detailed Sample Preparation Procedure to Produce Samples for Hierarchical Cluster Analysis

Dried *Angelica keiskei* Koidzumi root material was acquired from Strictly Medicinal Seeds® in Williams, Oregon, and a voucher specimen was deposited at the UNC Herbarium at Chapel Hill (NCU627665). Fresh *Angelica keiskei* Koidzumi roots were dried in a single-wall transite oven (Blue M Electric Company, Blue Island, IL, USA) at 40°C for 24 hours, producing 138.90 g of dry material. This material was ground using a Wiley Mill Standard Model No. 3 (Arthur Thomas Co., Philadelphia, PA, USA) and submerged in MeOH for 24 hours at 160 g/L. Plant material was filtered from extract and resuspended in equal volume of methanol. This process was repeated over three days. The resulting MeOH extract was concentrated *in vacuo* and subjected to liquid-liquid partitioning. First, defatting was completed by partitioning 10% aqueous MeOH and hexane (1:1). The aqueous MeOH layer was partitioned again between 4:5:1 EtOAc/MeOH/H_2O. Finally, to remove hydrosoluble tannins, the EtOAc layer was washed with a 1% NaCl aqueous solution (1:1). The resulting EtOAc extract (3,650.32 mg) was dried under nitrogen before further experimentation.

The EtOAc crude extract was subjected to a 40 minute round of flash chromatography using a Combiflash RF instrument (Teledyne ISCO, Lincoln, NE, USA). The gradient was held at 100% hexane for 3 min, ramped up to 100% chloroform over 20 min, and held at 100% chloroform for 9 min. Over the next three min, the gradient was increased to 20% methanol and 80% chloroform and held for five min, following which it was increased to 100% methanol over two min. Finally, the gradient was held at 100%

methanol for one minute. The extract was divided into nine pools. The ninth pool was collected from 20 to 100% methanol, and is the subject of the remaining experimentation.

The ninth pool (126.4 mg) was combined with four known compounds: alpha-mangostin (1.66 mg, 1% total mass), cryptotanshinone (3.32 mg, 2% total mass), magnolol (11.63 mg, 7% total mass), and berberine (24.92 mg, 15% of pool mass). These compounds were added to the mixture to enable evaluation of the effectiveness of our filtering approach and subsequent statistical analyses using a mixture of known and unknown compounds at varying concentrations.

Protocol S2: Plant Extraction and Simplification of *Angelica keiskei* Fraction

Plant material and extraction

Fresh *Angelica keiskei* roots were collected on November 14, 2015 in Williams, Oregon from Strictly Medicinal Seeds ® (Sample # 12444, N 42°12'17.211", W 123°19'34.60"). The identity of the sample was confirmed by Richard A. Cech and a voucher specimen was deposited at the University of North Carolina Chapel Hill Herbarium (NCU627665). Fresh root material was dried at 40°C for 24 hours in a single-wall transite oven (Blue M Electric Company, Blue Island, IL, USA), yielding 138.9 g of dried root material. Roots were then ground to a powder using a Wiley Mill Standard Model No. 3 (Arthur Thomas Col, Philadelphia, PA, USA). Powdered root was submerged in MeOH at 160 g/L for 24 hours, then filtered from the solvent. This process was repeated using the same root material every 24 hours for 72 hours. The resulting methanol extract was then subjected to liquid-liquid partitioning. Fats were separated from the mixture by partitioning 10% aqueous methanol and hexane 1:1). The aqueous/methanol layer was partitioned again using EtOAc/MeOH/H_2O (4:5:1). Lastly, hydrosoluble tannins were separated from the EtOAc layer by washing it with a 1% NaCl aqueous solution (1:1). The resulting EtOAc extract was dried under nitrogen, yielding 3,650.32 mg of material.

Production of simplified *A. keiskei* fraction

The EtOAc extract was separated using a 40 min normal-phase gradient conducted on a Combiflash RF instrument (Teledyne ISCO, Lincoln, NE, USA). The gradient began with a 3 min hold at 100% hexane, after which it was increased to 100%

chloroform over the next 20 min. It was then held at 100% chloroform for 9 min, after which the gradient was increased to 20:80 MeOH:$CHCl_3$ over 3 min. These conditions were held for five min, after which it was increased to 100% methanol over two min. The gradient was held at 100% methanol for one min. The resulting tubes were separated into nine fractions and subjected to biological activity testing. The ninth fraction was collected from 20-100% methanol, and was used for the remainder of the experimental procedures, due to its lack of antimicrobial activity (<15% inhibition at 100 µg/mL against a laboratory strain of *Staphylococcus aureus*, SA1199).

Protocol S3: Chromatographic Separation and Isolation of *Salvia miltiorrhiza*

The first-stage separations of the EtOAc extract (SM) were conducted on an aliquot of 8.6 g of the extract using normal-stage flash chromatography (120-g silica column) at an 85 mL/min flow rate with a 45-min hexane/CH_3Cl/MeOH gradient. Two fractions, SM-1 and SM-3, were selected for further chromatographic separation. The first fraction (SM-1, 185.72 mg) was subjected to reversed-phase preparative HPLC injected onto a Gemini preparatory column (5 µm C18, 250 x 21.20 mm; Phenomenex) at a flow rate of 21.4 mL/min with a 45-min gradient. The gradient began at 65:35 CH_3CN:H_2O and increased to 90:10 over 35 min, following which the column was held at 100:0 for 10 min, yielding 8 fractions. Fraction 5 (SM-1-5, 36.51 mg) was subjected to a final round of reversed-phase preparative HPLC injected onto a Gemini preparatory column (5 µm C18, 250 x 21.20 mm; Phenomenex). The 30 min run began at 70:30 CH_3CN:H_2O and was increased to 100:0 over 30 min. Compound **5** (SM-1-5-5) eluted from 12-14 min (1.39 mg, 98% purity, 0.0003% yield). Fraction SM-3 (1058.67 mg) was subjected to a second round of normal-phase flash chromatography (40-g silica solumn) at a flow rate of 40 mL/min and a 55 min hexane/CH_3Cl/MeOH gradient, yielding four fractions. Fraction one (SM-3-1, 844.33 mg) eluted from 6-9 min, and was subjected to an additional round of reversed-phase flash chromatography using an 86g C18 reversed-phase RediSep Rf column with a 60 mL/min flow rate. A 60-min gradient of CH_3CN was used ranging from 45-100% CH_3CN. Compound **1** eluted at 25 min (580.01 mg, 95.0% purity, 0.1% yield).

Compound **4** was isolated using the remaining 9.7 g of the EtOAc extract (SM). First, normal-stage flash chromatography (80-g silica column) was conducted with a 40-min hexane/CH_3Cl/MeOH gradient and a 60 mL/min flow rate, yielding 8 fractions (SM-9 through SM-16). The fourth fraction, SM-12 (391.90 mg), was subjected to a second round of flash chromatography (12-g silica column, 30 mL/min) separated using a 45 gradient of hexane/EtOAc/MeOH. Of the seven resulting fractions (SM-12-1 through SM-12-7), the fourth fraction, SM-12-4 (108.01 mg), was fractionated using reversed-phase HPLC. The sample was injected onto a Gemini preparatory column (5 µm C18, 250 x 21.20 mm; Phenomenex) at a flow rate of 21.4 mL/min with a 45-min gradient. The gradient began at 40:60 CH3CN:H2O and increased to 50:50 over 35 min, after which the column was increased to 100:0 and held for 10 min, yielding 7 fractions (SM-12-4-1 through SM-12-4-7). Fraction SM-12-4-5 (3.19 mg) was purified with a final round of reversed-phase chromatography using a Gemini semi-preparatory column (5 µm C18, 250 x 10.00 mm; Phenomenex) at a flow rate of 4.7 mL/min and a 45-min gradient ranging from 43-48% CH_3CN. Compound **4** eluted at 18 min (0.5 mg, 93% purity, 0.0001% yield).

APPENDIX B

SUPPLEMENTARY TABLES

Table S1. Complete List of Chemical Contaminants Removed from Analysis using Hierarchical Cluster Analysis Coupled to Spectral Variable Inspection of Triplicate Injections.

Table S2. Effect of Data Acquisition Protocols on Selectivity Ratio Analyses.

Table S3. False Positives and their Distribution in Selectivity Ratio Models.

Table S4. Effect of Data Processing Protocols on Selectivity Ratio Analyses.

Table S5. Effect of Round of Fractionation on Selectivity Ratio Analyses.

Table S6. Comparison of Stage-One Models and their Identification of Randainal among the Top Contributors to Biological Activity.

Table S7. Complete List of Chemical Contaminants Removed from Analysis using Hierarchical Cluster Analysis Coupled to Spectral Variable Inspection of Triplicate Injections from *S. miltiorrhiza* Extracts.

Table S8. NMR Data for Sugiol (Compound 5) in $CDCl_3$.

Table S1. Complete List of Chemical Contaminants Removed from Analysis using Hierarchical Cluster Analysis Coupled to Spectral Variable Inspection of Triplicate Injections. Chemical contaminants were consistent across samples.

Accurate Mass	Retention Time (min)	Tentative Identification*	Ion type	Found in MeOH Blank?	Found with quantitative filter? [a]
215.094	3.837			Y	Y
217.049	6.534			N	Y
265.147	7.021			Y	Y
265.149	7.192			Y	Y
281.048	8.472			Y [β]	Y
297.154	6.95			Y	Y
355.07	7.795			Y	Y
445.121	7.116			Y	Y
503.108	8.477	Polysiloxane, $[C_2H_6SiO]_7$	$[M+H-CH_4]^+$	Y	Y
504.105	8.473	Polysiloxane, $[C_2H_6SiO]_7$	$[M+H-CH_4]^+$, ^{13}C isotope	Y	Y
504.11 [δ]	8.493			Y	Y
505.106	8.477	Polysiloxane, $[C_2H_6SiO]_7$	$[M+H-CH_4]^+$, 2 × ^{13}C isotope	Y	Y
519.139	8.474	Polysiloxane, $[C_2H_6SiO]_7$	$[M+H]^+$	Y [β]	Y
520.139	8.473	Polysiloxane, $[C_2H_6SiO]_7$	$[M+H]^+$, ^{13}C isotope	Y	Y
521.118	8.583			Y	N
521.136	8.474	Polysiloxane, $[C_2H_6SiO]_7$	$[M+H]^+$, 2 × ^{13}C isotope	Y [β]	Y
522.136 [δ]	8.487			Y [β]	N
522.147 [δ]	7.647			Y [β]	Y
522.153 [δ]	7.629			Y [β]	Y
523.115 [δ]	8.589			Y	N
523.15 [δ]	7.634			Y [β]	Y
524.115 [δ]	8.585			Y	Y
524.127 [δ]	8.501			Y	N
524.144 [δ]	7.636			Y [β]	Y
524.15 [δ]	7.636			Y [β]	Y
525.147 [δ]	7.634			Y	Y
536.166	6.688	Polysiloxane, $[C_2H_6SiO]_7$	$[M+NH_4]^+$	Y	Y
536.166	8.472	Polysiloxane, $[C_2H_6SiO]_7$	$[M+NH_4]^+$	Y	Y
537.163	8.469	Polysiloxane, $[C_2H_6SiO]_7$	$[M+NH_4]^+$, ^{13}C isotope	Y	Y
537.168	8.496	Polysiloxane, $[C_2H_6SiO]_7$	$[M+NH_4]^+$, ^{13}C isotope	Y	Y
538.144 [δ]	8.488			N	Y
538.162	8.479	Polysiloxane, $[C_2H_6SiO]_7$	$[M+NH_4]^+$, 2 × ^{13}C isotope	Y	Y

538.168	8.525	Polysiloxane, $[C_2H_6SiO]_7$	$[M+NH_4]^+$, 2 × ^{13}C isotope	Y	Y
539.145 [δ]	8.486			Y	Y
539.164 [δ]	8.472			Y	Y
540.144 [δ]	8.494			Y	Y
540.162 [δ]	8.472			Y	Y
541.116 [δ]	8.473			Y [β]	N
541.122 [δ]	8.476			Y	N
541.157 [δ]	8.468			Y	Y
541.163 [δ]	8.469			Y	Y
542.12 [δ]	8.472			Y	N
542.156 [δ]	8.471			Y [β]	N
542.162 [δ]	8.472			Y	N
550.182	8.475			Y	Y
557.094	8.472			Y	Y
564.195	8.473			Y	Y
582.151	8.776			Y	Y
610.186	7.161	Polysiloxane, $[C_2H_6SiO]_8$	$[M+NH_4]^+$	Y	Y
611.181	7.437	Polysiloxane, $[C_2H_6SiO]_8$	$[M+NH_4]^+$, ^{13}C isotope	Y	Y
611.188	7.154	Polysiloxane, $[C_2H_6SiO]_8$	$[M+NH_4]^+$, ^{13}C isotope	Y	Y
611.188	7.83	Polysiloxane, $[C_2H_6SiO]_8$	$[M+NH_4]^+$, ^{13}C isotope	Y	Y
612.185	7.158	Polysiloxane, $[C_2H_6SiO]_8$	$[M+NH_4]^+$, 2 × ^{13}C isotope	Y	Y
612.186	7.171	Polysiloxane, $[C_2H_6SiO]_8$	$[M+NH_4]^+$, 2 × ^{13}C isotope	Y	Y
613.185 [δ]	7.158			Y	Y
613.185 [δ]	7.714			Y	N
614.18 [δ]	7.16			Y [β]	Y
670.185	8.997			Y	Y
671.189	8.997			Y	Y
684.198 [δ]	8.247			N	Y
684.206	8.03	Polysiloxane, $[C_2H_6SiO]_9$	$[M+NH_4]^+$	Y	Y
684.206	8.63	Polysiloxane, $[C_2H_6SiO]_9$	$[M+NH_4]^+$	Y	N
685.2	8.024	Polysiloxane, $[C_2H_6SiO]_9$	$[M+NH_4]^+$, ^{13}C isotope	Y	Y
685.208	8.029	Polysiloxane, $[C_2H_6SiO]_9$	$[M+NH_4]^+$, ^{13}C isotope	Y	Y
686.179 [δ]	7.827			Y [β]	Y
686.187 [δ]	7.778			Y	Y
686.197 [δ]	7.866			Y [β]	Y
686.197 [δ]	8.155			Y [β]	Y
686.205	8.018	Polysiloxane, $[C_2H_6SiO]_9$	$[M+NH_4]^+$, 2 × ^{13}C isotope	Y [β]	Y
686.214 [δ]	8.659			Y	N
686.222 [δ]	8.67			Y [β]	N
687.195 [δ]	7.885			Y [β]	Y

687.203 [δ]	8.013			Y	Y
688.195 [δ]	7.901			Y [β]	Y
688.203 [δ]	8.653			Y	Y
744.201	8.67			Y [β]	Y
744.211	8.695			Y [β]	Y
745.204	8.69			Y	Y
746.188	8.663			Y [β]	N
746.198	8.668			Y	N
746.208	8.678			Y [β]	N
747.185	8.666			Y [β]	N
747.194	8.67			Y	N
747.204	8.671			Y [β]	N
748.183	8.66			Y	N
748.193	8.664			Y [β]	N
748.203	8.671			Y [β]	N
749.184	8.661			Y	N
758.221	8.378	Polysiloxane, $[C_2H_6SiO]_{10}$	$[M+NH_4]^+$	Y	Y
759.222	8.377	Polysiloxane, $[C_2H_6SiO]_{10}$	$[M+NH_4]^+$, ^{13}C isotope	Y	Y
760.204 [δ]	8.936			Y	Y
760.215 [δ]	8.394			Y [β]	Y
760.225	8.369	Polysiloxane, $[C_2H_6SiO]_{10}$	$[M+NH_4]^+$, 2 × ^{13}C isotope	Y	Y
760.235 [δ]	8.372			Y [β]	Y
761.199 [δ]	8.947			Y	Y
761.22 [δ]	8.379			Y	Y
761.23 [δ]	8.373			Y [β]	Y
761.24 [δ]	8.368			Y [β]	Y
762.197 [δ]	8.955			Y	Y
762.207 [δ]	8.962			Y	Y
762.217 [δ]	8.377			Y	Y
762.227 [δ]	8.371			Y	Y
762.237 [δ]	8.372			Y [β]	N
763.216 [δ]	8.374			Y [β]	Y
795.167	5.398			N	N
818.222	8.592			Y	Y
819.222	8.62			Y	N
834.215	8.964			Y [β]	Y
834.224	8.957			Y	Y
834.236	8.987			Y	Y
834.246	8.989			N	Y
835.218	8.965			Y [β]	Y
835.23	8.962			Y	Y
835.241	8.995			Y [β]	Y
836.215	8.964			Y [β]	N
836.226	8.964			Y [β]	Y
836.238	8.973			Y [β]	Y
836.25	8.955			N	N
837.214	8.964			Y	N
837.225	8.963			Y	N
837.238	8.969			Y [β]	N

838.214	8.912	Y	N
906.263	8.948	Y[β]	Y
907.26	8.939	Y[β]	Y
907.261	8.597	Y	Y
908.246	8.407	Y[β]	Y
908.259	8.981	Y[β]	Y
909.26	8.998	Y	Y

* Tentative identifications accomplished using Interferences and Contaminants Encountered in Modern Mass Spectrometry, Keller et al. (271)

[a] Spectral variables receiving a "Y" in this category had an average variance/mean peak area within triplicate injections greater than 1.0×10^7 if found using a low-concentration dataset, or 4.1×10^7 if using a high-concentration dataset. Those receiving an "N" in this category were only identified using visual inspection of chromatograms

[β] These *m/z* retention time pairs were found in some, but not all, of the blank injections

[δ] These masses represent peaks we believe to be associated with polysiloxane isotopes (containing more than $2 \times {}^{13}C$) and/or mass spectral artefacts. They were too low abundant to be fragmented using the LC-MS data analysis method, so they could not be confirmed to be the same as tentatively identified polysiloxanes. Instead, we have tentatively identified them by their similarity in accurate mass/retention time to putatively identified polysiloxanes from Keller et al. (271)

Table S2. **Effect of Data Acquisition Protocols on Selectivity Ratio Analyses.** We assessed the impact of pool number, bioassay concentration, and mass spectral concentration on final biochemometric results by evaluating changes in the selectivity ratio ranking of berberine and magnolol, as well as the impact on false postives identified in the models.

Subset	# Fractions	Conc. tested in bioassay (ug/mL)	Conc. analyzed in MS (mg/mL)	Number of ions included in model (*m/z* / RT pairs)	Model Produced? (Y/N)	Number of model components	% independent, % dependent	SR ranking berberine	SR ranking magnolol	# false positive co-varying with berberine	# false positives co-varying with magnolol	Number of false positive not co-varying
1[a]	3	100	0.1	870	Y	4	99.99, 99.92	1	20	2	16	1
2	3	50	0.1	870	N	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3	3	25	0.1	870	Y	5	99.99, 99.95	N/A	14	0	17	0
4	5	100	0.1	870	Y	2	99.38, 84.98	1	14	1	15	0
5	5	50	0.1	870	Y	2	99.37, 86.40	1	12	1	15	0
6	5	25	0.1	870	Y	2	99.38, 84.82	1	14	1	15	0
7	10	100	0.1	870	Y	5	99.79, 98.55	1	8	2	22	0
8	10	50	0.1	870	Y	5	99.79, 82.00	22	4	0	22	0
9	10	25	0.1	870	Y	5	99.81, 88.07	1	13	2	25	8
10	3	100	0.01	370	Y	5	99.98, 100	7	27	0	18	7
11	3	50	0.01	370	N	N/A	N/A	N/A	N/A	N/A	N/A	N/A
12	3	25	0.01	370	N	N/A	N/A	N/A	N/A	N/A	N/A	N/A
13	5	100	0.01	370	Y	4	99.71, 99.83	7	20	0	19	4
14	5	50	0.01	370	Y	3	99.57, 99.76	20	16	0	19	4
15	5	25	0.01	370	Y	3	99.57, 99.73	1	20	0	19	4
16	10	100	0.01	370	N	N/A	N/A	N/A	N/A	N/A	N/A	N/A
17	10	50	0.01	370	Y	2	94.96, 49.17	33	18	0	30	11
18	10	25	0.01	370	Y	3	62.28, 79.11	1	36	0	28	28

[a]Cryptotanshinone correctly identified as contributing to activity (19th). Cryptotanshinone only contributed to activity in 3 pool set.

Table S3. False Positives and their Distribution in Selectivity Ratio Models.

# Fractions	Concentration tested in bioassay	Concentration analyzed in MS	Number of ions included in model[a]	Number of ions with selectivity ratio > 0 (% total[b])	% associated with berberine and magnolol[c]	% co-varying false positives[c]	% non-co-varying false positives[c]
3	100	0.1	870	26 (3%)	27%	69%	4%
3	50	0.1	--	--	--	--	--
3	25	0.1	870	20 (2%)	15%	85%	0%
5	100	0.1	870	22 (3%)	28%	72%	0%
5	50	0.1	870	22 (3%)	28%	72%	0%
5	25	0.1	870	22 (3%)	28%	72%	0%
10	100	0.1	870	30 (3%)	20%	80%	0%
10	50	0.1	870	28 (3%)	21%	79%	0%
10	25	0.1	870	41 (5%)	34%	66%	0%
3	100	0.01	370	33 (9%)	25%	55%	20%
3	50	0.01	--	--	--	--	--
3	25	0.01	--	--	--	--	--
5	100	0.01	370	32 (9%)	28%	59%	13%
5	50	0.01	370	32 (9%)	28%	59%	13%
5	25	0.01	370	32 (9%)	28%	59%	13%
10	100	0.01	--	--	--	--	--
10	50	0.01	370	50 (14%)	18%	60%	22%
10	25	0.01	370	65 (18%)	14%	43%	43%

[a] representing unique *m/z* / RT pairs

[b] expressed as a percentage of the total number of ions included in model

[c] expressed as a percentage of the total number of ions with selectivity ratio > 0.

Table S4. **Effect of Data Processing Protocols on Selectivity Ratio Analyses.** All models contained 870 unique mass/retention time pairs and were produced using data acquired from the 10-pool set analyzed at 100 µg/mL in both the biological assay and during mass spectral analysis.

Data Transformation?	Dendrogram Filtering?	Percent Variance Cutoff?	Number of model component	% independent, % dependent	SR ranking berberine	SR ranking magnolol	# false positives co-varying with berberine[a]	Number of false positives co-varying with magnolol[a]	Number of contaminants identified with dendrogram analysis in model[a,b]	Number of false positive not co-varying[a]
N	N	N	5	99.77, 98.71	23	120	13	0	4	27
N	N	Y	5	99.77, 98.71	2	9	3	21	1[c]	0
N	Y	N	5	99.79, 98.55	17	110	20	1	N/A	25
N	Y	Y	5	99.79, 98.55	1	8	2	22	N/A	0
Y	N	N	5	79.90, 99.77	17	213	17	3	2	21
Y	N	Y	5	79.90, 99.77	17	205	17	3	2	21
Y	Y	N	5	81.10, 99.75	19	200	18	3	N/A	22
Y	Y	Y	5	81.10, 99.75	19	192	19	3	N/A	21

[a] Only top 50 ions were included in this summary

[b] These contaminants were identified and removed using dendrogram filtering, so models that went through dendrogram filtering will not have this type of contaminant in the model

[c] polysiloxane contaminant peak identified as top contributor to bioactivity

Table S5. Effect of Round of Fractionation on Selectivity Ratio Analyses.

Round of Fractionation	# Fractions	Concentration tested in bioassay (ug/mL)	Model Produced? (Y/N)	Number of model components	% independent, % dependent	SR ranking magnolol	# false positives co-varying with magnolol [a]	Number of false positive not co-varying [a]
1	3	50	N	N/A	N/A	N/A	N/A	N/A
2	11	50	Y	1	32.62, 86.52	1	18	0
1	3	25	Y	5	99.99, 99.95	14	17	0
2	11	25	Y	1	31.39, 88.97	6	18	0
1	5	50	Y	2	99.37, 86.40	12	13	0
2	10	50	Y	1	43.68, 91.27	1	15	1
1	5	25	Y	2	99.38, 84.82	14	13	0
2	10	25	Y	1	42.97, 72.03	2	16	1
1	10	50	Y	5	99.79, 82.00	4	18	0
2	7	50	Y	2	61.92, 94.10	N/A	6	12 [b]
1	10	25	Y	5	99.81, 88.07	13	10	2
2	7	25	Y	1	36.95, 76.91	4	16	0

[a] only top twenty contributors were considered for this metric

[b] in this case, an unexpected active compound (randainal) was identified as the fifth top contributor to activity. Likely, the activity of this compound was masked by antagonists until this round of fractionation. Nine of the 12 "non-co-varying false positives" actually co-varied with randainal, and only 3 represented actual false positives that did not co-vary with an active compound.

Table S6. Comparison of Stage-One Models and their Identification of Randainal among the Top Contributors to Biological Activity.

Subset	Round of Fractionation	# Fractions	Conc. tested in bioassay (µg/mL)	Conc. analyzed in MS (mg/mL)	Model Produced? (Y/N)	Number of model components	% independent, % dependent	Did model identify randainal?	SR ranking of randainal
1	1	3	100	0.1	Y	4	99.99, 99.92	Y	23
2	1	3	50	0.1	N	N/A	N/A	N/A	N/A
3	1	3	25	0.1	Y	5	99.99, 99.95	Y	17
4	1	5	100	0.1	Y	2	99.38, 84.98	N	N/A
5	1	5	50	0.1	Y	2	99.37, 86.40	N	N/A
6	1	5	25	0.1	Y	2	99.38, 84.82	N	N/A
7	1	10	100	0.1	Y	5	99.79, 98.55	Y	19
8	1	10	50	0.1	Y	5	99.79, 82.00	Y	14
9	1	10	25	0.1	Y	5	99.81, 88.07	Y	25
10	1	3	100	0.01	Y	5	99.98, 100	N	N/A
11	1	3	50	0.01	N	N/A	N/A	N/A	N/A
12	1	3	25	0.01	N	N/A	N/A	N/A	N/A
13	1	5	100	0.01	Y	4	99.71, 99.83	N	N/A
14	1	5	50	0.01	Y	3	99.57, 99.76	N	N/A
15	1	5	25	0.01	Y	3	99.57, 99.73	N	N/A
16	1	10	100	0.01	N	N/A	N/A	N/A	N/A
17	1	10	50	0.01	Y	2	94.96, 49.17	N	N/A
18	1	10	25	0.01	Y	3	62.28, 79.11	N	N/A
19	2	11	50	0.1	Y	1	32.62, 86.52	Y	50
20	2	11	25	0.1	Y	1	31.39, 88.97	Y	49
21	2	10	50	0.1	Y	1	43.68, 91.27	N	N/A
22	2	10	25	0.1	Y	1	42.97, 72.03	N	N/A
23	2	7	50	0.1	Y	2	61.92, 94.10	Y	5
24	2	7	25	0.1	Y	2	62.68, 86.41	N	N/A

Table S7. Complete List of Chemical Contaminants Removed from Analysis using Hierarchical Cluster Analysis Coupled to Spectral Variable Inspection of Triplicate Injections in *S. miltiorrhiza* Samples. Chemical contaminants were consistent across samples.

Accurate Mass	Ionization Mode	Retention Time (min)	Tentative Identification*	Ion Type
279.159	Positive	8.661	Dibutylphthalate	$[M+H]^+$
336.636	Positive	5.879		
357.133	Negative	3.873		
357.133	Positive	4.04		
357.134	Negative	4.295		
357.134	Positive	4.696		
367.117	Positive	4.383		
536.166	Positive	8.548	Polysiloxane, $[C_2H_6SiO]_7$	$[M+NH_4]^+$
537.166	Positive	8.553	Polysiloxane, $[C_2H_6SiO]_7$	$[M+NH_4]^+$, ^{13}C isotope
537.147†	Positive	8.558		
538.165	Positive	8.558	Polysiloxane, $[C_2H_6SiO]_7$	$[M+NH_4]^+$, $2 \times {}^{13}C$ isotope
539.149†	Positive	8.558		
539.165†	Positive	8.551		
539.208	Positive	7.308		
540.161†	Positive	8.551		
541.161†	Positive	8.549		
837.216	Positive	8.97		
837.224	Positive	8.57		

* Tentative identifications accomplished using Interferences and Contaminants Encountered in Modern Mass Spectrometry, Keller et al. (271)

† These masses represent peaks we believe to be associated with polysiloxane isotopes (containing more than $2 \times {}^{13}C$) and/or mass spectral artefacts. They were too low abundant to be fragmented using the LC-MS data analysis method, so they could not be confirmed to be the same as tentatively identified polysiloxanes. Instead, we have tentatively identified them by their similarity in accurate mass/retention time to putatively identified polysiloxanes from Keller et al. (271)

Table S8. NMR Data for Sugiol (Compound 5) in $CDCl_3$. ^{1}H, HMBC, and HSQC data collected at 500 MHz, and ^{13}C data collected at 125 MHz

Position	**^{13}C**	**^{1}H**	**HMBC**
1α	37.97*	1.53 m*	
1β		2.23 dt (J=11.9, 2.8)	
2α	18.97	1.67 m	
2β		1.76 tt (J=13.6, 3.3)	
3α	41.42	1.25 m*	
3β		1.53 m*	
4	33.37		
5	49.53	1.85 dd (J=13.7, 4.0)	9
6α	36.13	2.68 dd (J=18.1, 4.0)	5
6β		2.58 dd (J=18.1, 13.8)	
7	198.68		
8	124.78		
9	156.52		
10	37.95*		
11	110.03	6.68 s	10, 8, 13, 12
12	158.15		
13	132.63		
14	126.63	7.90 s	15, 9, 12, 7
15	26.88	3.12 hept (J=6.9)	
16	22.55	1.24 d (J=6.9)	13, 15, 17
17	22.42	1.26 d (J=6.9)	13, 16, 15
18	32.65	0.92 s	19, 3, 5
19	21.45	0.98 s	18, 3, 5
20	23.33	1.21 s	1, 5, 9

* overlapping assignments based on HSQC experiments.

APPENDIX C

SUPPLEMENTARY FIGURES

Figure S37: ^{1}H NMR Data for 1-oxocryptotanshinone (500 MHz, $CDCl_3$)

Figure S38. Calibration Curve of Cryptotanshinone used to Quantify Cryptotanshinone in each *S. milttiorhiza* Fraction.

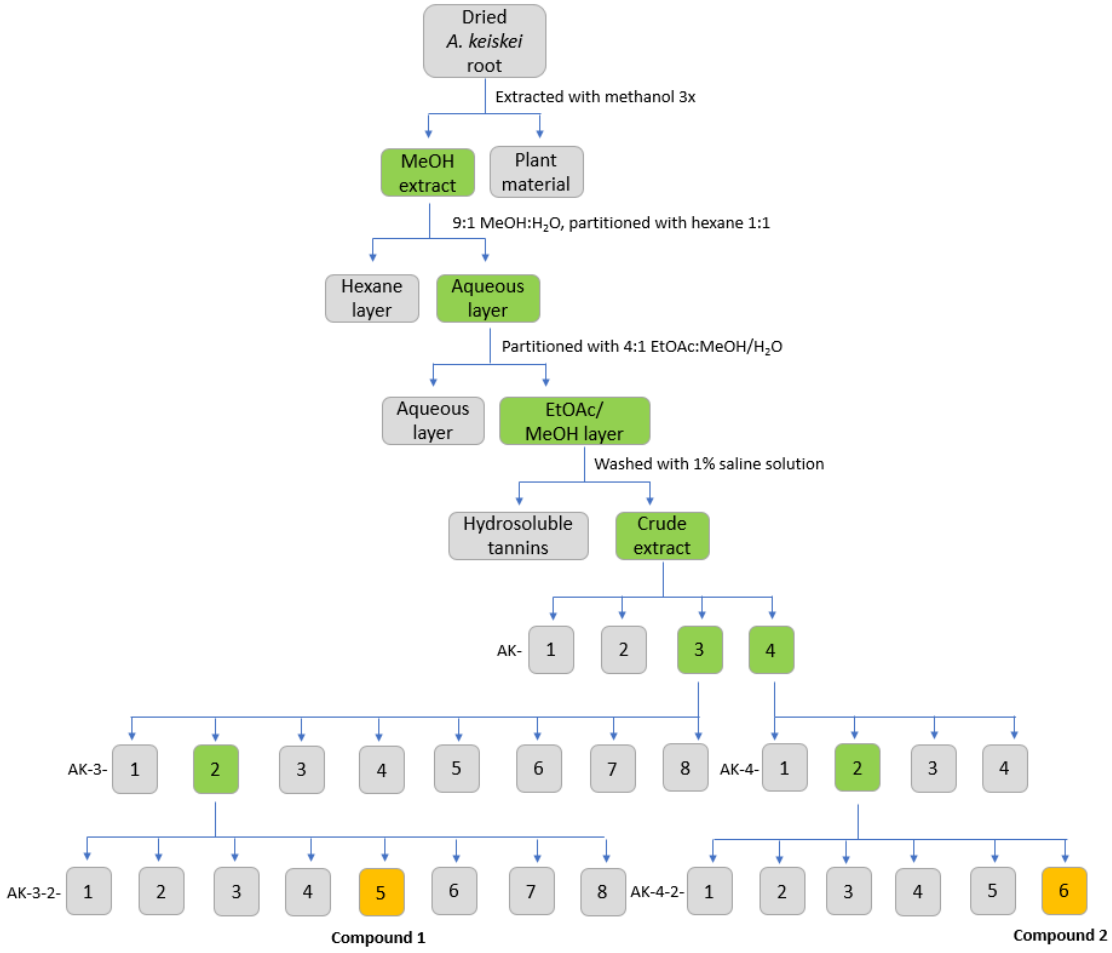


Figure S1. Fractionation Scheme: *Angelica keiskei* Molecular Networking and Biochemometrics.

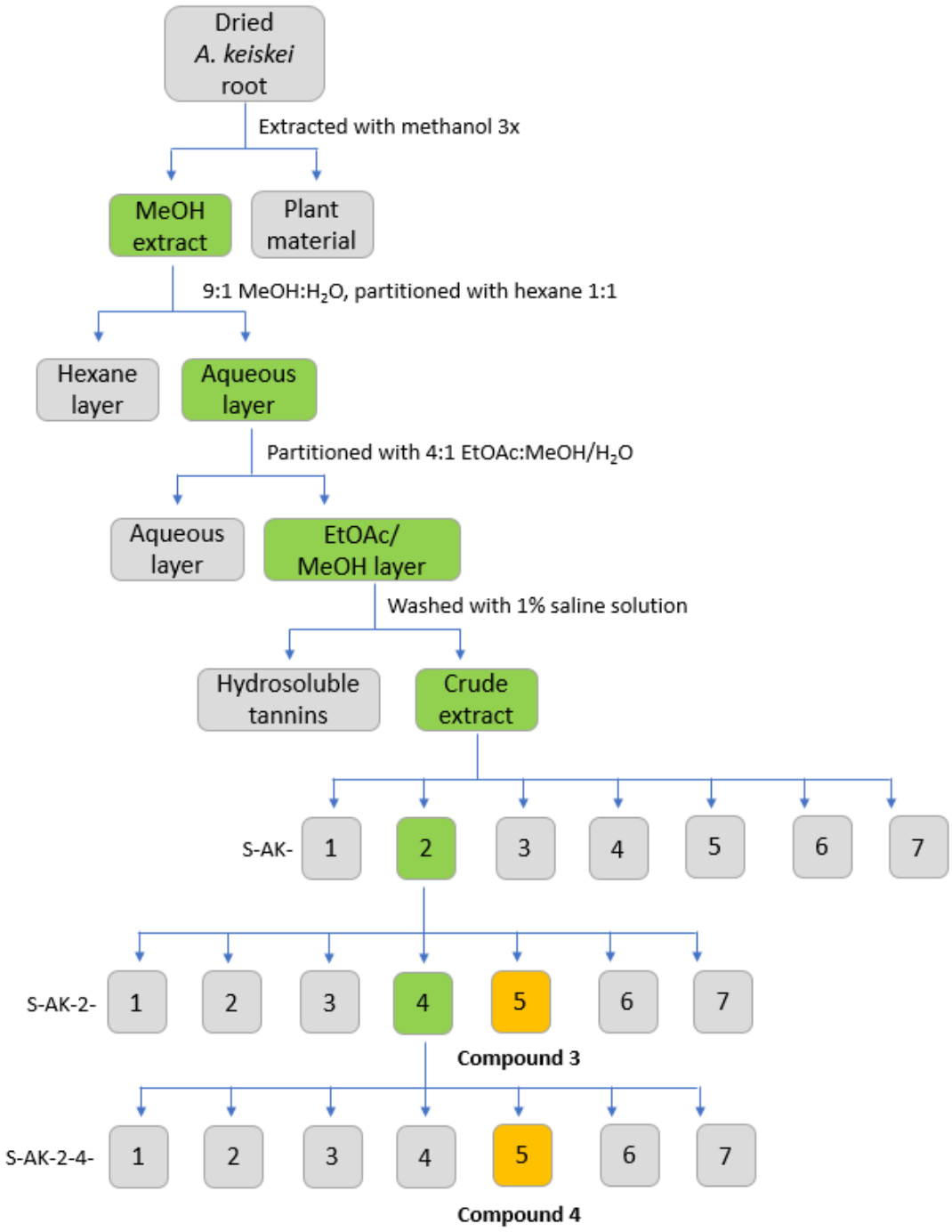


Figure S2. Scale-up Fractionation Scheme: *Angelica keiskei.*

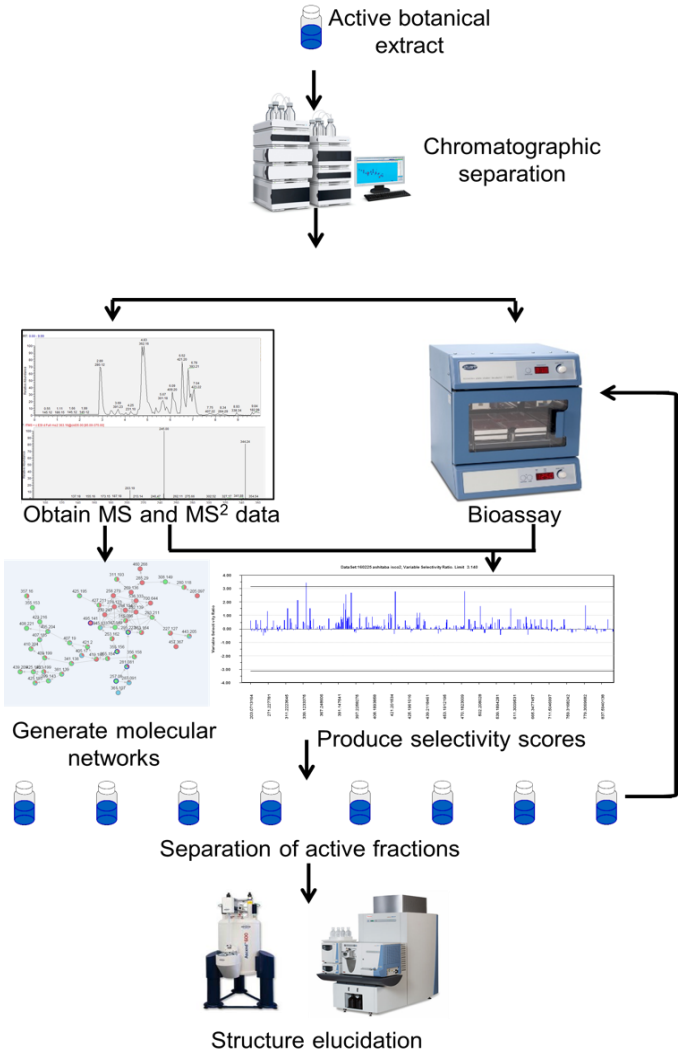


Figure S3. Analytical Workflow for Integrated Analysis.

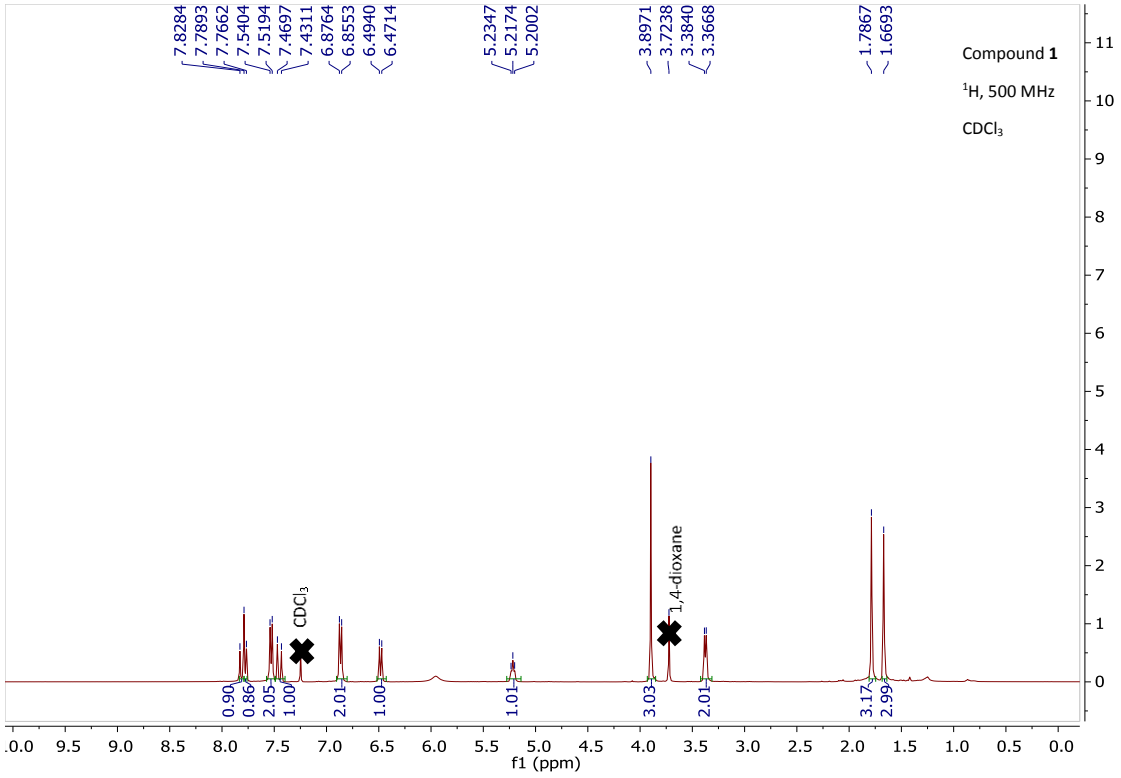


Figure S4. ^{1}H NMR Spectrum of (500 MHz, $CDCl_3$) 4-hydroxyderricin.

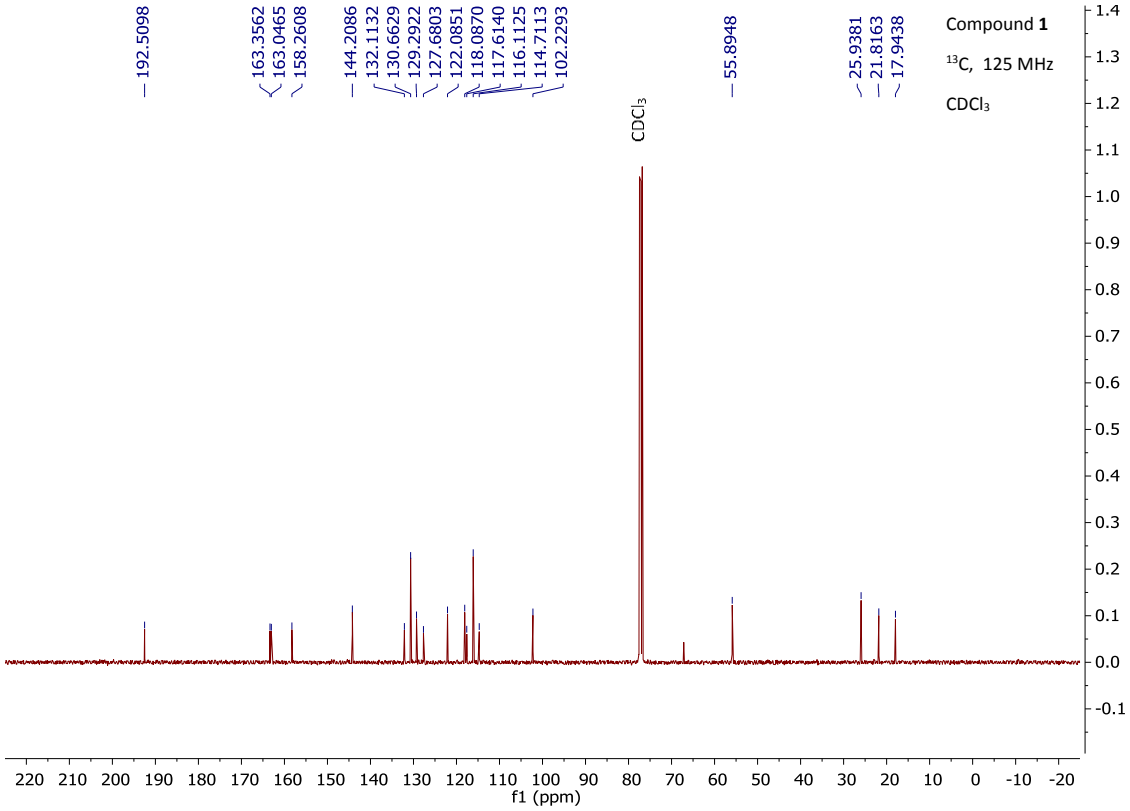


Figure S5. ^{13}C NMR Spectrum (125 MHz, $CDCl_3$) of 4-hydroxyderricin.

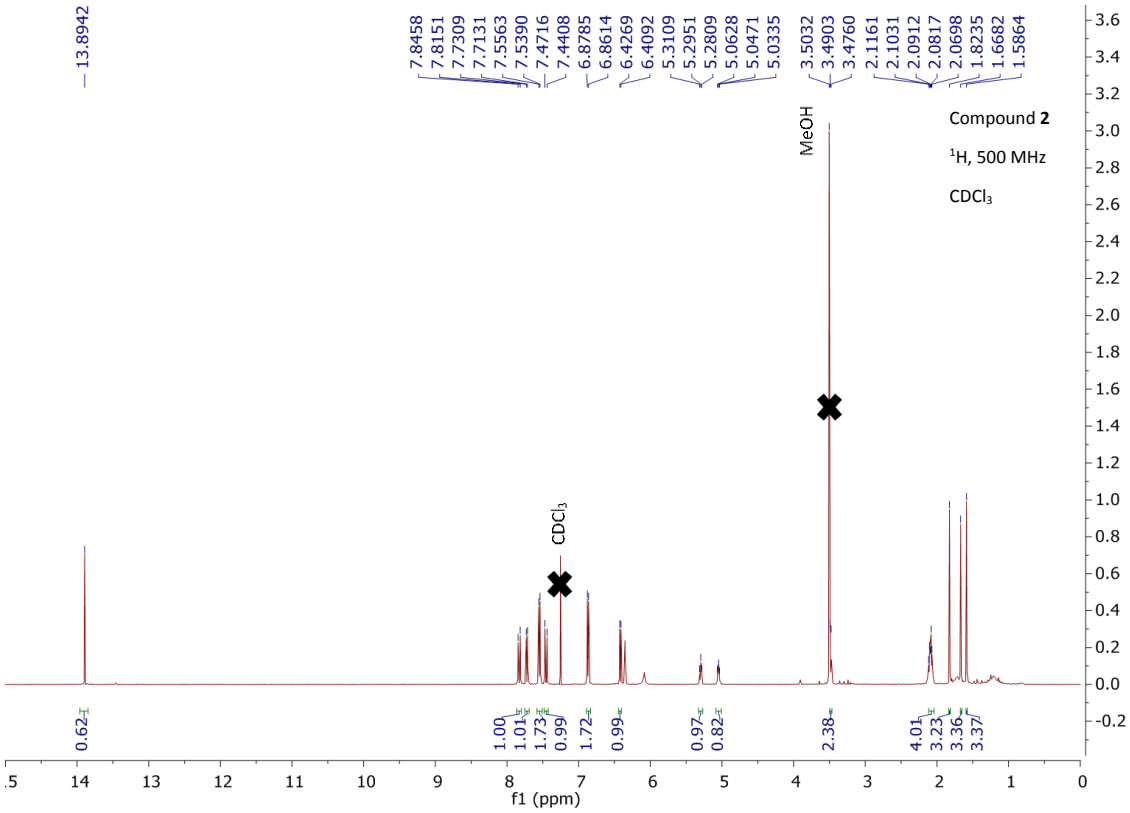


Figure S6. ^{1}H NMR Spectrum (500 MHz, $CDCl_3$) of Xanthoangelol.

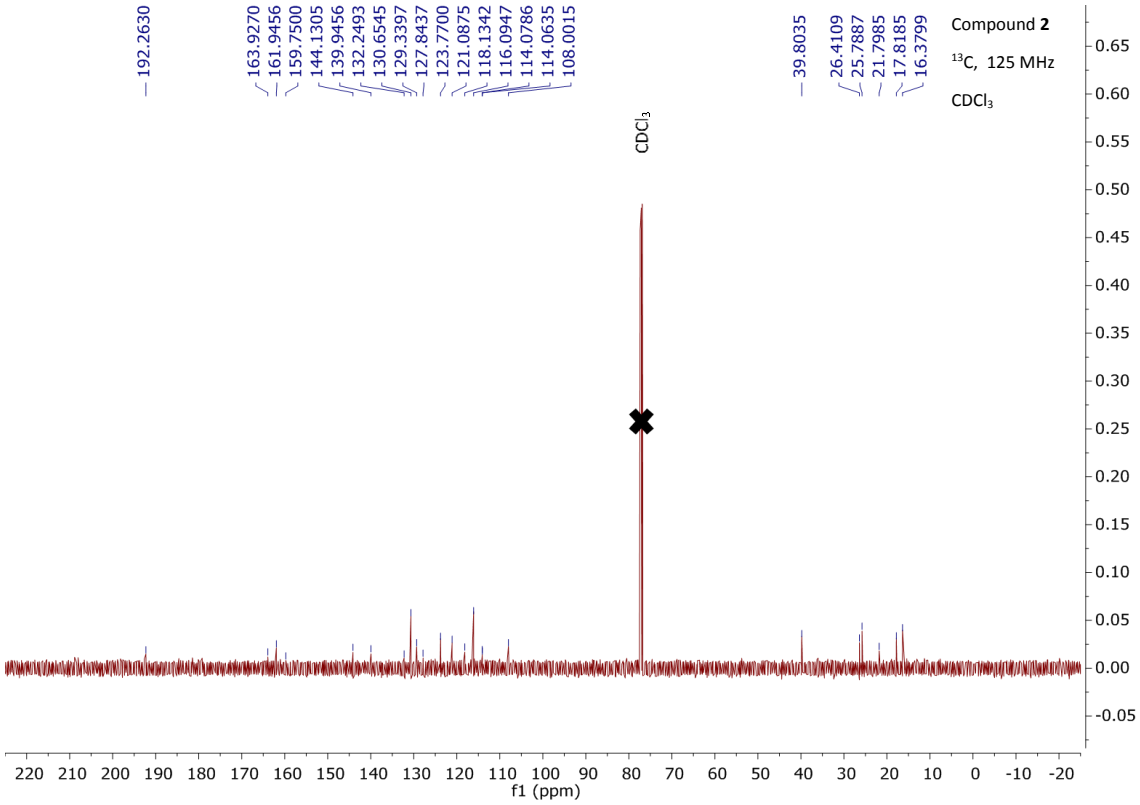


Figure S7. ^{13}C NMR Spectrum (125 MHz, $CDCl_3$) of Xanthoangelol.

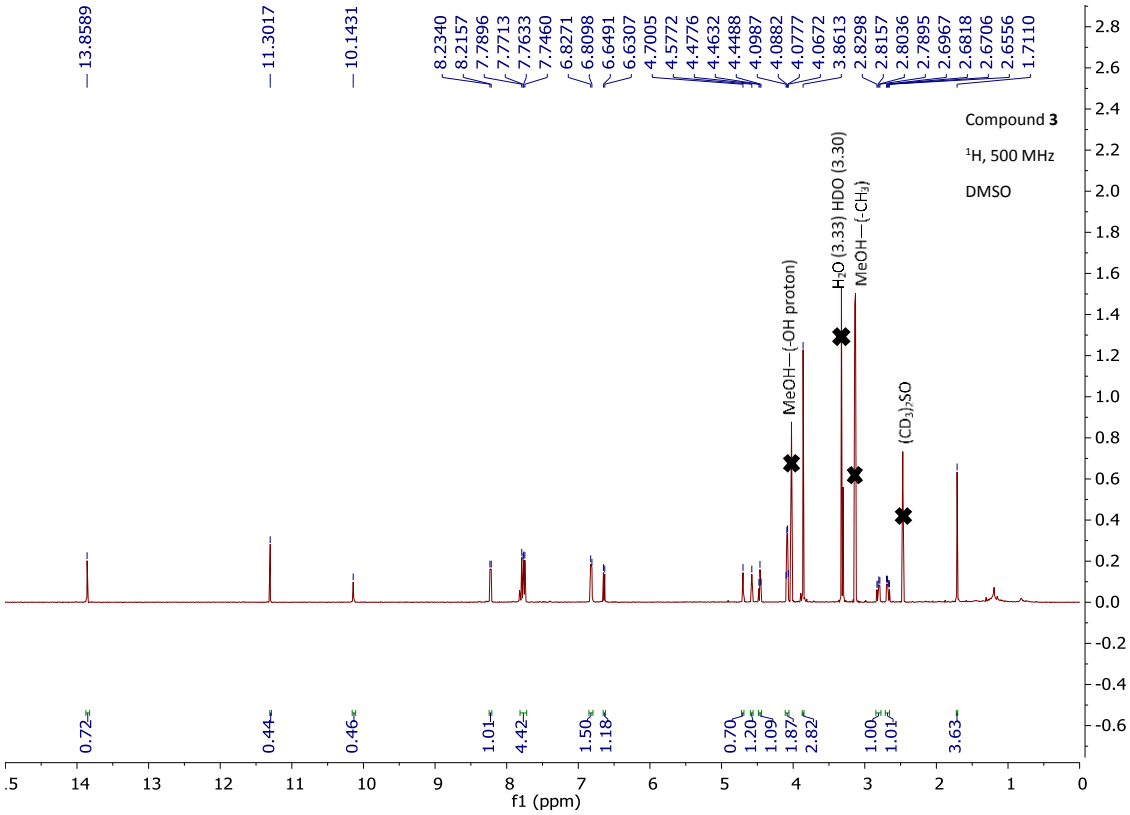


Figure S8. ^{1}H NMR Spectrum (500 MHz, DMSO-*d6*) of Xanthoangelol E.

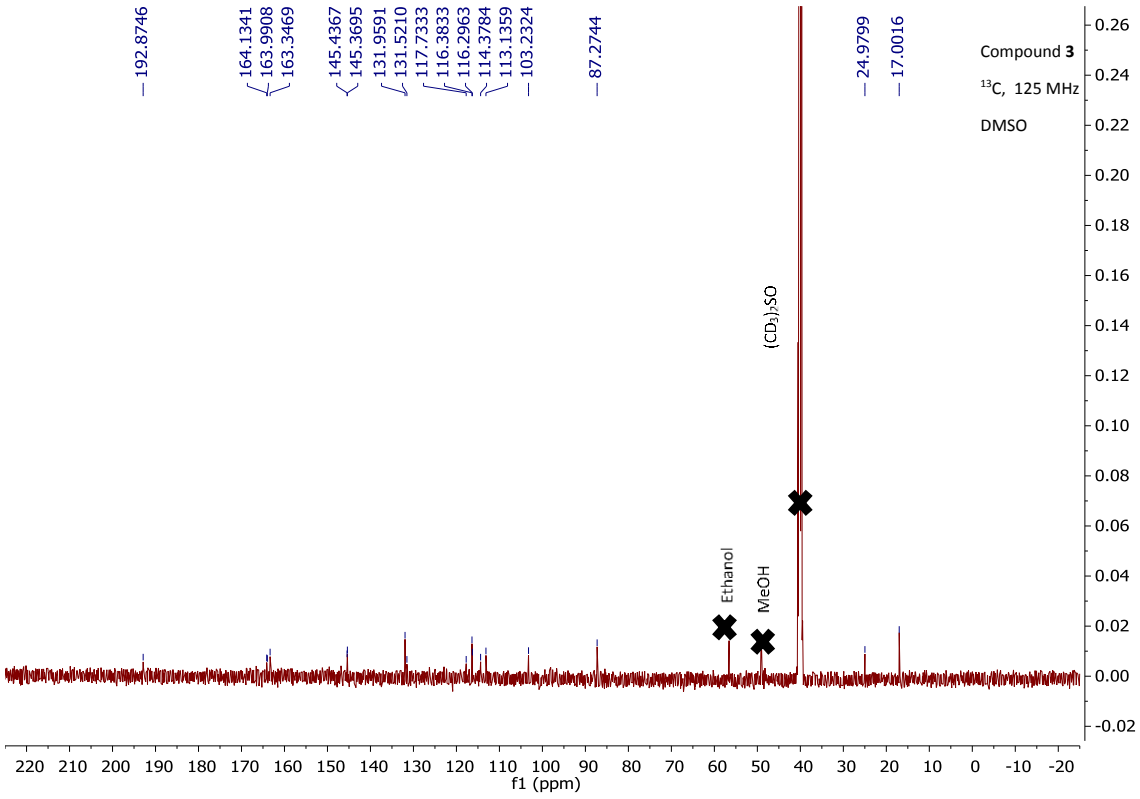


Figure S9. ^{13}C NMR Spectrum (125 MHz, DMSO-*d*$_6$) of Xanthoangelol E.

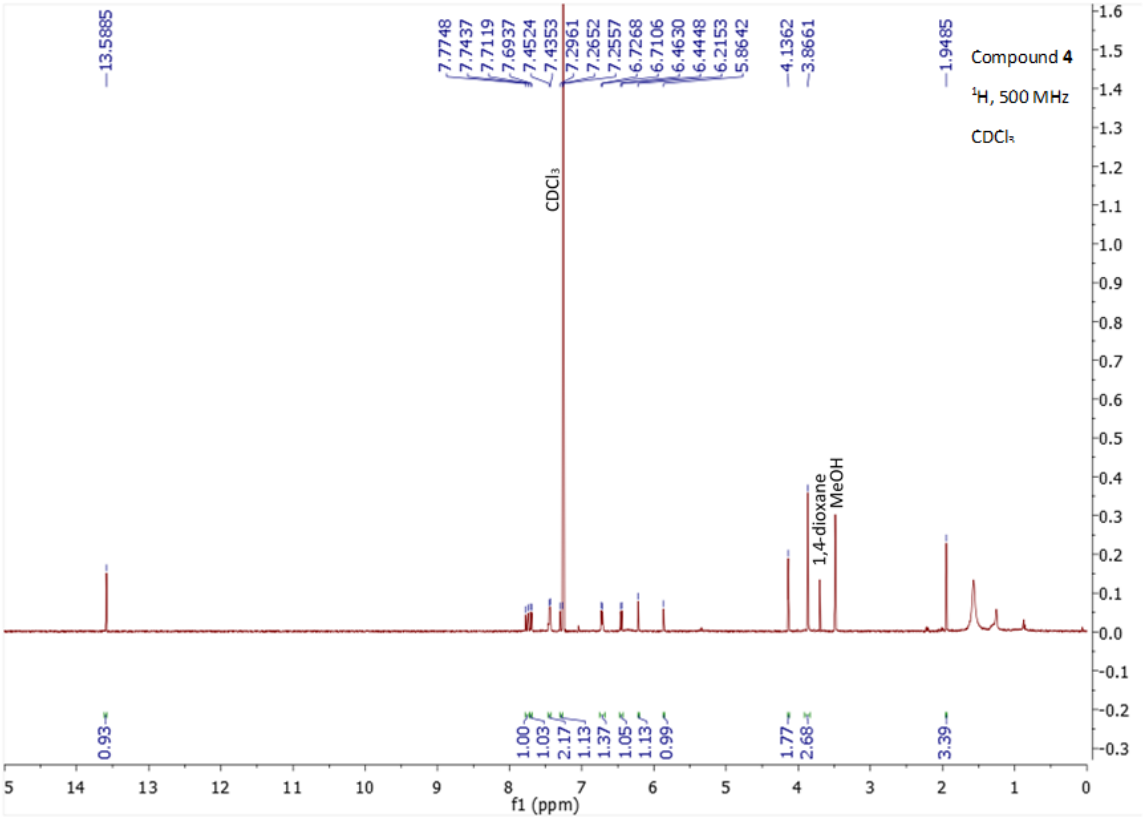


Figure S10. ^{1}H NMR Spectrum (500 MHz, $CDCl_3$) of Xanthoangelol K.

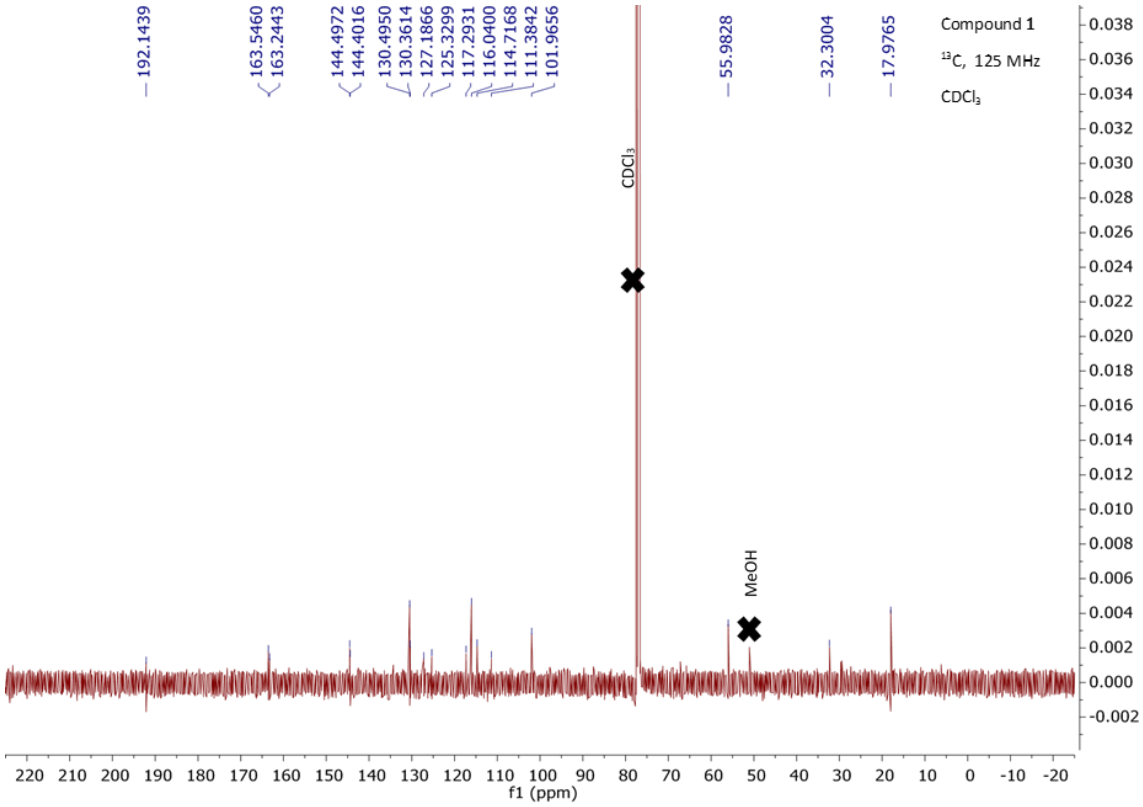


Figure S11. ^{13}C NMR Spectrum (125 MHz, $CDCl_3$) of Xanthoangelol K.

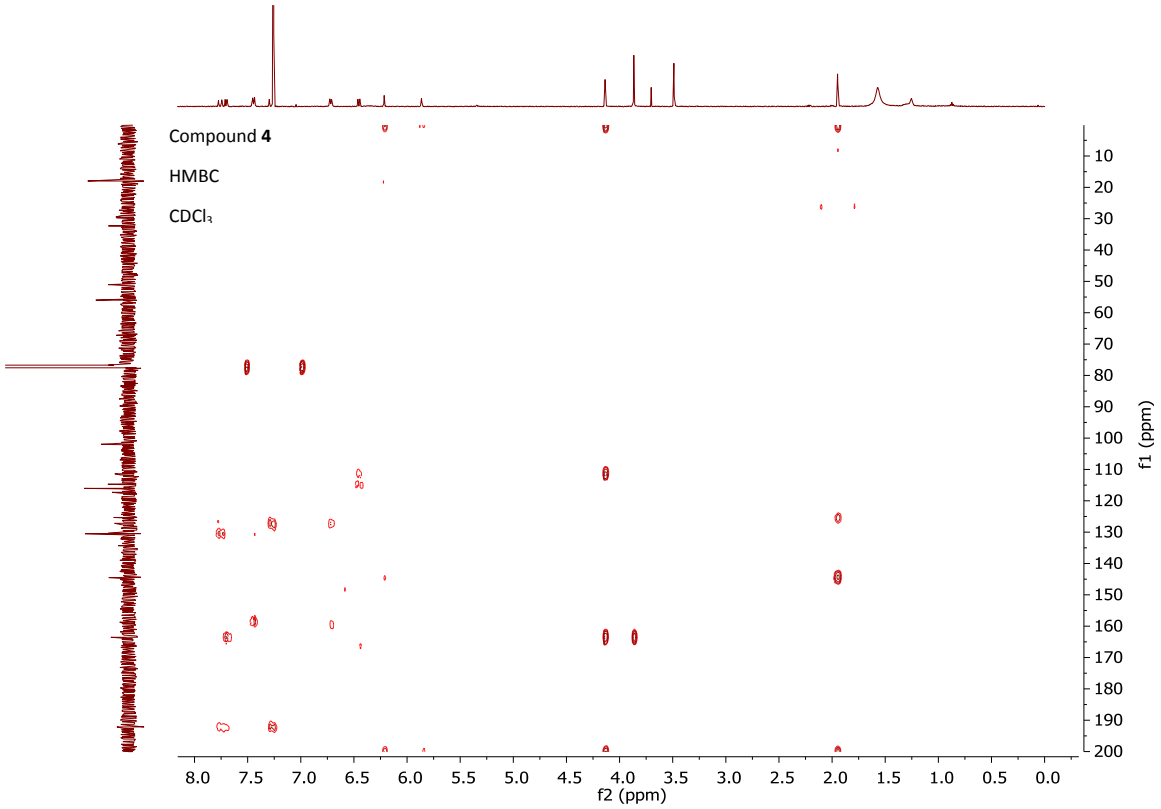


Figure S12. HMBC Spectrum (400 MHz, $CDCl_3$) of Xanthoangelol K.

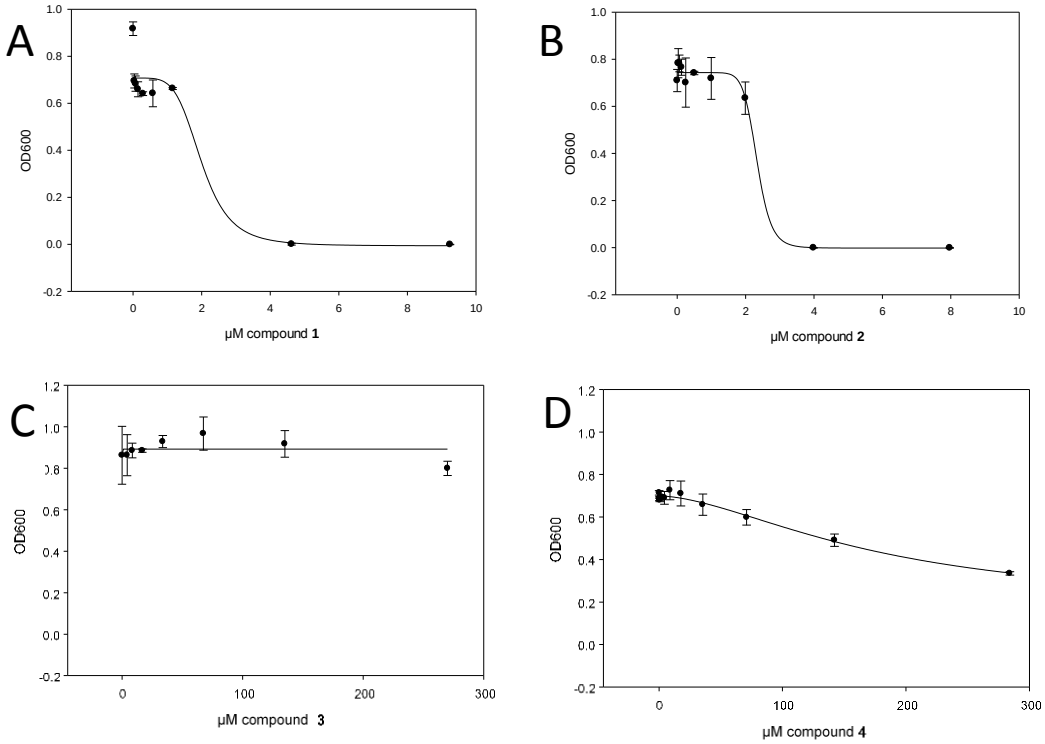


Figure S13. Dose Response Curves for 4-hydroxyderricin (A), Xanthoangelol (B), Xanthoangelol E (C), and Xanthoangelol K (D) Isolated from *A. keiskei* against MRSA USA300 LAC Strain AH1263 (234). Turbidimetric data were obtained by comparing OD_{600} values of test wells relative to vehicle control after 24 hours of incubation at 37 °C. Models were constructed using untransformed triplicate data and fit using four-parameter logistic curves, represented as the mean ± SEM.

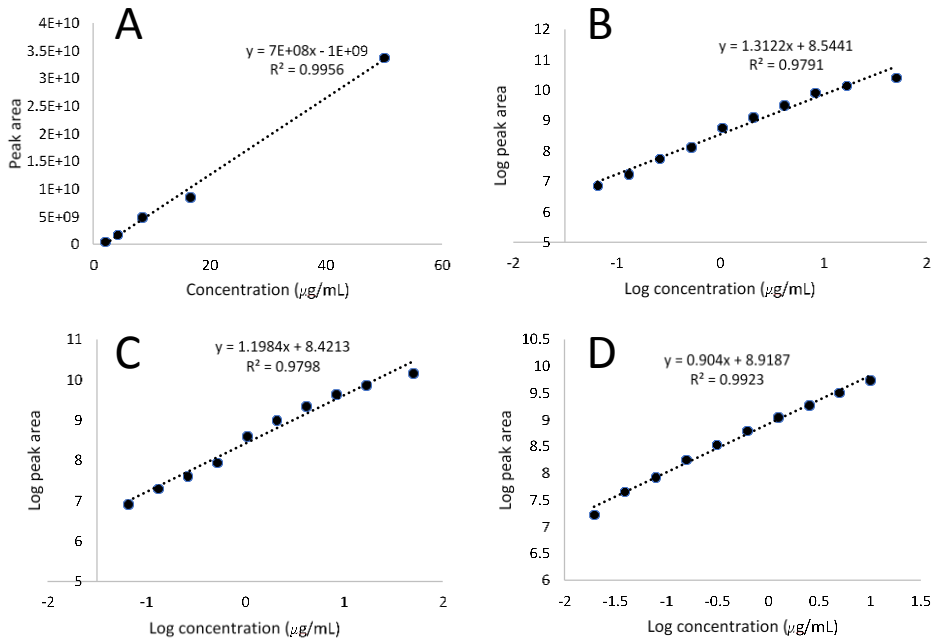


Figure S14. Calibration Curves of Standard Compounds of Berberine (A), Magnolol (B), Cryptotanshinone (C), and Alpha-mangostin (D). Curves were produced using a Thermo-Fisher Q-Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific, MA, USA) connected to an Acquity UPLC system (Waters Corporation, Milford, MA, USA). Separations were completed by using a reversed phase UPLC column (BEH C18, 1.7 µm, 2.1 x 50 mm, Waters Corporation, Milford, MA, USA).

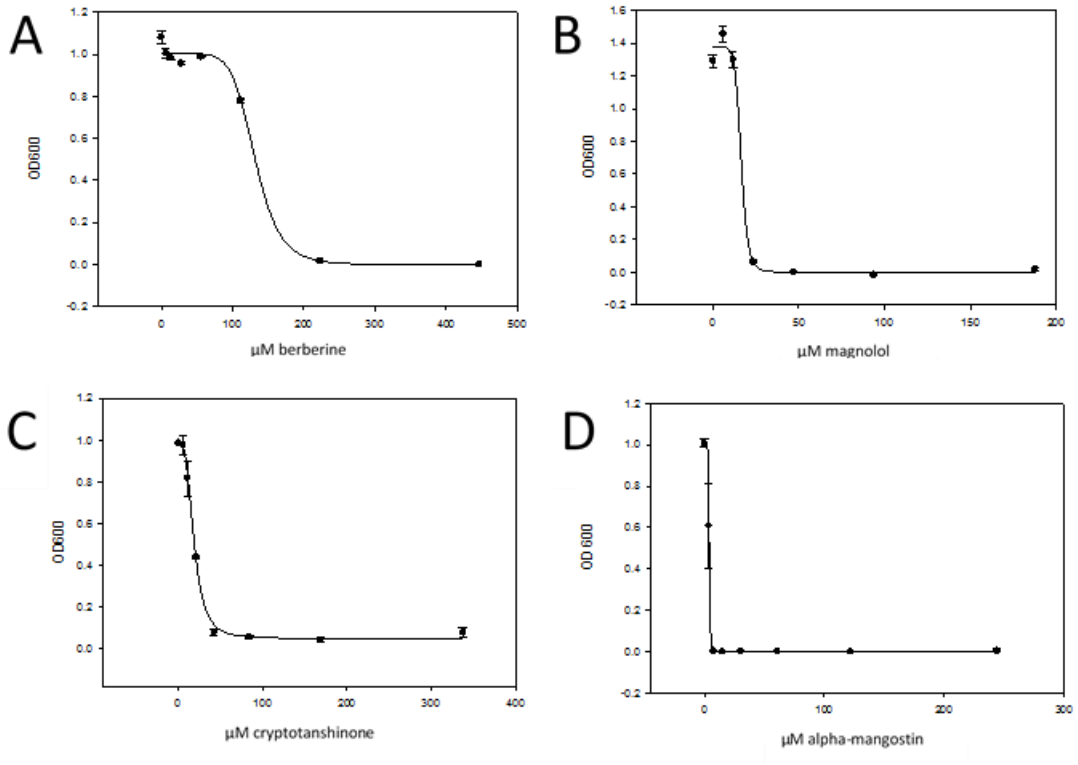


Figure S15. Dose-Response Curves of Berberine (A), Magnolol (B), Cryptotanshinone (C), and Alpha-mangostin (D) against *S. aureus* SA1199. Turbidimetric data were obtained by comparing OD_{600} values of test wells relative to vehicle control following 18 hours of incubation at 37 °C. Models were constructed using untransformed triplicate data and fit using four-parameter logistic curves, represented as the mean ± SEM.

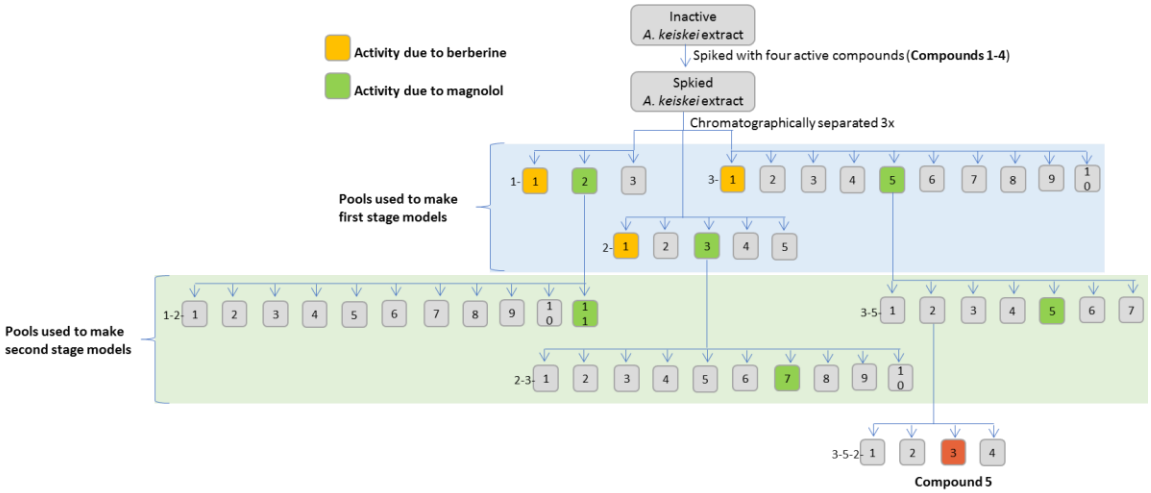


Figure S16. **Fractionation Scheme.** Pools used to produce first- and second-stage models have been identified in brackets.

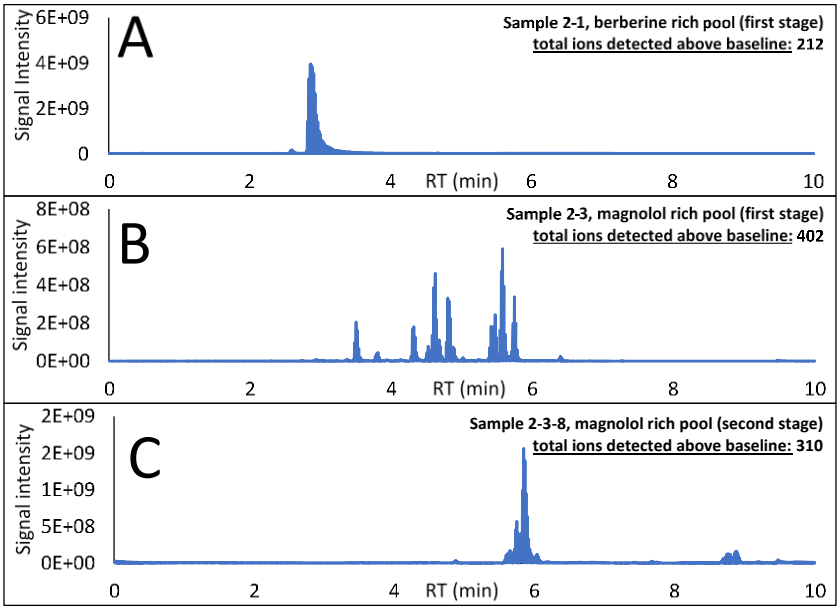


Figure S17. **Example Chromatograms of Active Fractions from First- and Second-Stages of Fractionation.** Data produced using both positive- and negative-mode data of selected active pools belonging to the 5-pool set analyzed at 0.1 mg/mL in the mass spectrometer. For first-stage pools, baseline cutoffs were set to 2.0×10^6 for positive mode and 1.0×10^6 for negative mode. For second-stage pools, baseline cutoffs were set to 2.0×10^6 for both positive mode and negative mode. **S4A**. Berberine-rich pool from the first round of chromatographic separation. **S4B**. Magnolol-rich pool from first round of chromatographic separation. **S4C**. Magnolol-rich pool from second round of chromatographic separation.

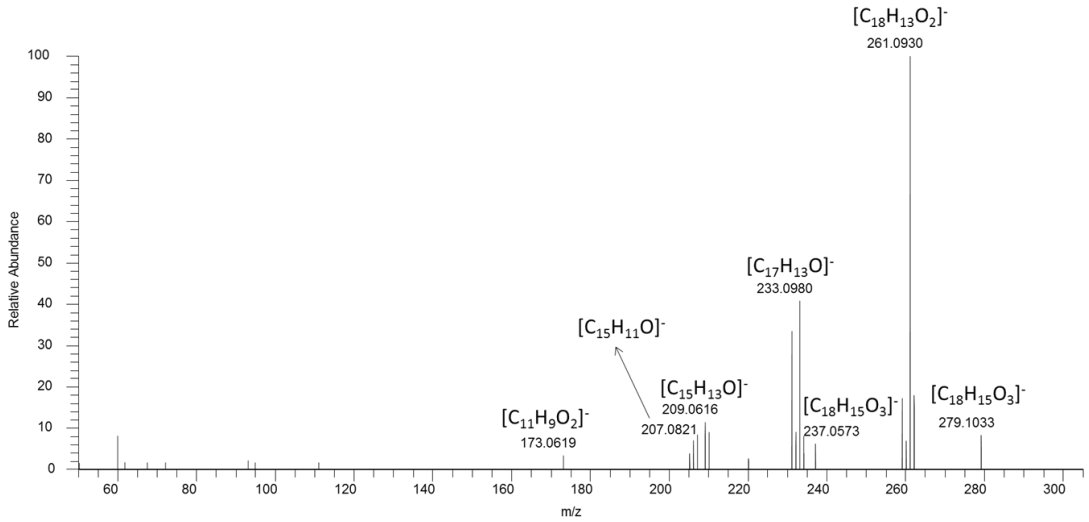


Figure S18. MS^2 Spectrum (Negative Mode) of Randainal. Peaks have been labeled with molecular formulas if they match fragment predictions and/or fragments previously reported in the literature (299).

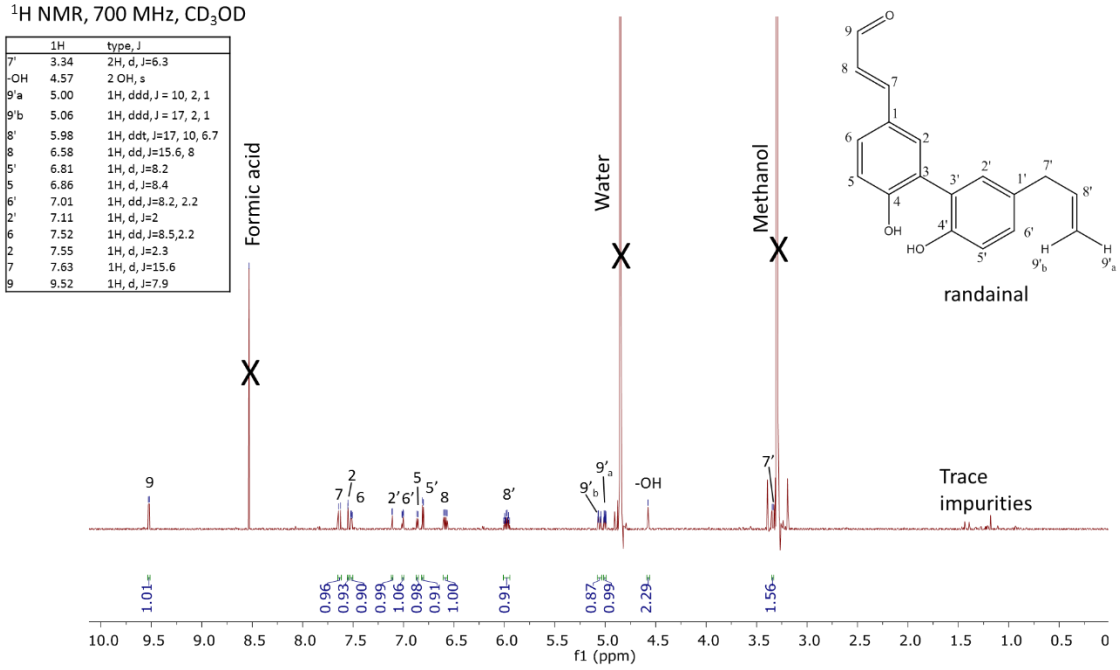


Figure S19. ^{1}H NMR Spectrum (700 MHz, CD_3OD) of Randainal.

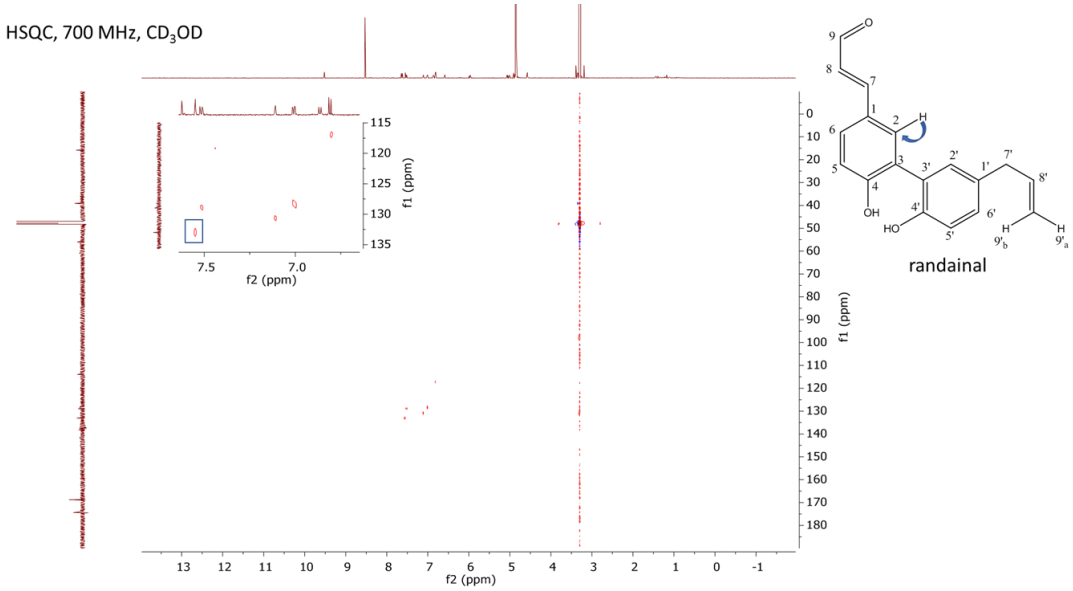


Figure S20. HSQC Spectrum (700 MHz, CD_3OD) of Randainal.

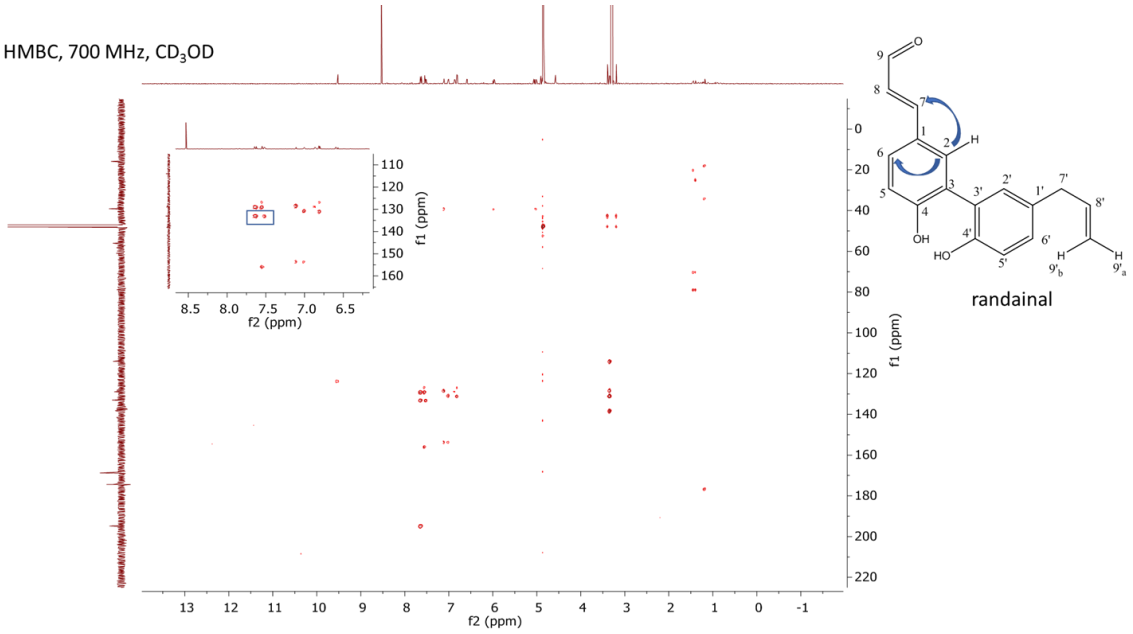


Figure S21. HMBC Spectrum (700 MHz, CD_3OD) of Randainal.

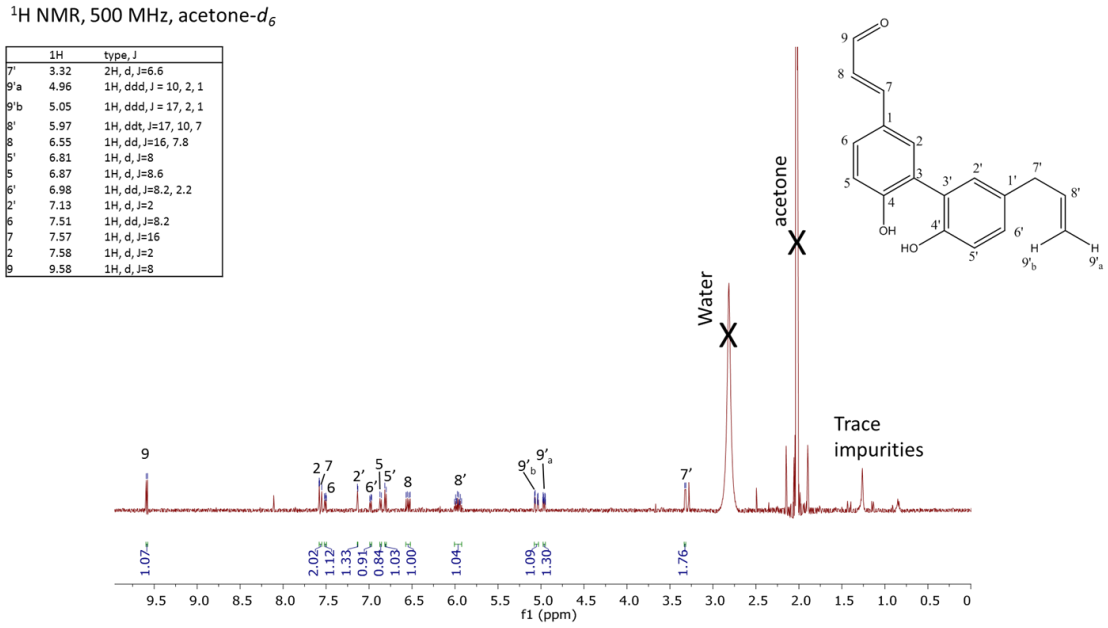


Figure S22. ^{1}H NMR Spectrum (500 MHz, Acetone-d_6) of Randainal. Spectra match those previously reported in the literature (299).

Figure S23. Fractionation Scheme for Simplify Development with *S. miltiorrhiza.*

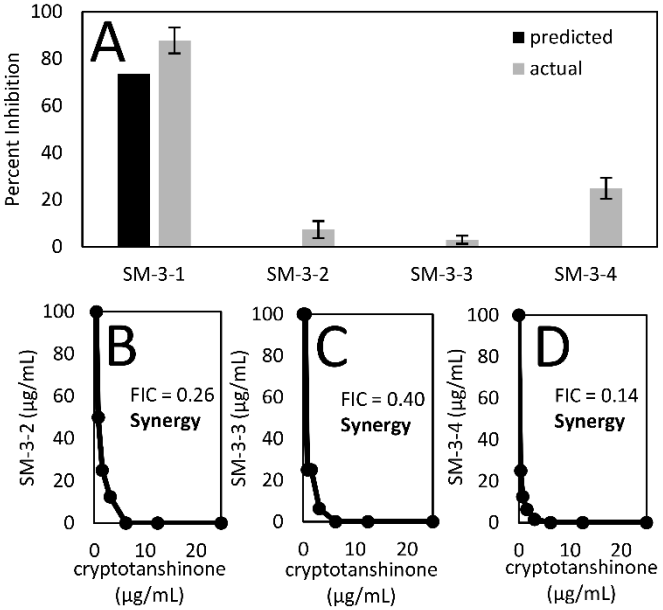


Figure S24A. Predicted versus Actual Activities of Sub-fractions Simplified from Synergistic Fraction SM-3. Although predicted and actual did not show a mismatch, we predicted that synergistic compounds were separated from cryptotanshinone which was used to calculate predicted activity. Cryptotanshinone was used as a positive control, and its MIC (25 µg/mL) is consistent with previous reports (285). Indeed, when isobolograms were generated for synergy testing, isobolograms of SM-3-2 (**B**), SM-3-3 (**C**), and SM-3-4 (**D**) all possessed synergy with FIC values of 0.26, 0.40, and 0.14 respectively.
ΣFICs were calculated using the following equation: $[A]/IC_{50}A + [B]/IC_{50}B = \Sigma FIC$, where $IC_{50}A$ is the IC_{50} of cryptotanshinone alone, $IC_{50}B$ is the IC_{50} of the fraction alone, [A] is the IC_{50} of cryptotanshinone in combination with fraction, and [B] is the IC_{50} of fraction in combination with cryptotanshinone. Synergy ≡ $\Sigma FIC < 0.5$, additivity ≡ $0.5 < \Sigma FIC < 1.0$, Indifference ≡ $1.0 < \Sigma FIC < 4.0$, Antagonism ≡ $\Sigma FIC > 4.0$

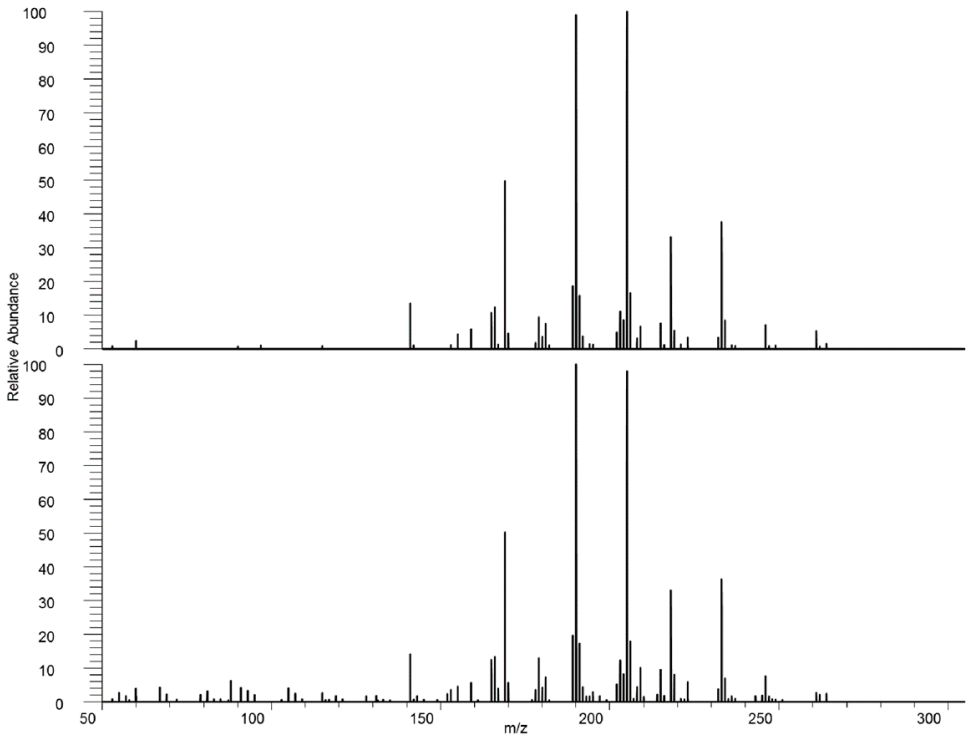


Figure S25. Fragmentation Patterns of Dihydrotanshinone I (HCD = 65). Fragmentation patterns of the pure standard compound (top) match fragmentation patterns of the compound found within the *S. miltiorrhiza* mixture (bottom).

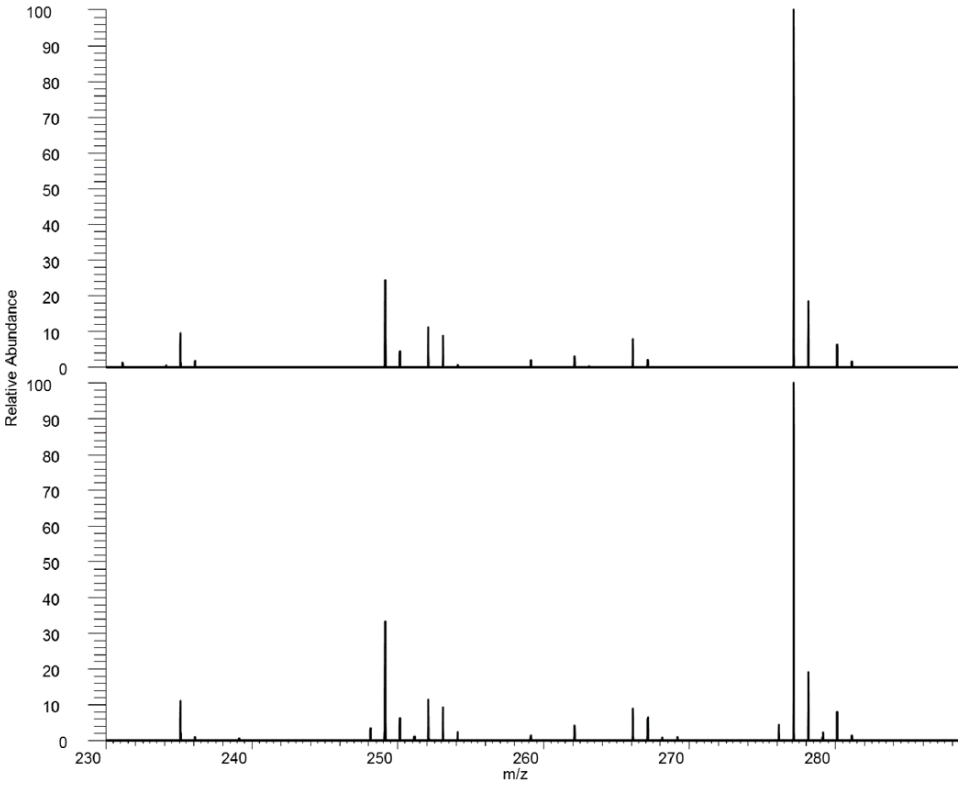


Figure S26. Fragmentation Patterns of Tanshinone IIA (HCD = 30). Fragmentation patterns of the pure standard compound (top) match fragmentation patterns of the compound found within the *S. miltiorrhiza* mixture (bottom).

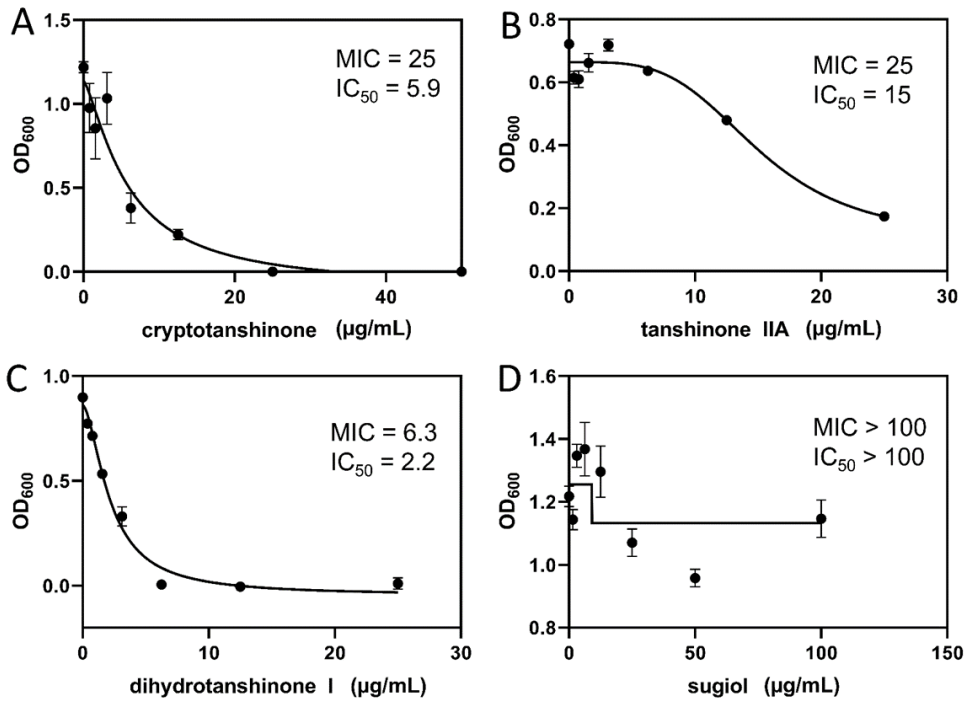


Figure S27. Dose Response Curves for Cryptotanshinone, Tanshinone IIA, Dihydrotanshinone I, and Sugiol. Curves were fit using a four-parameter logistic model.

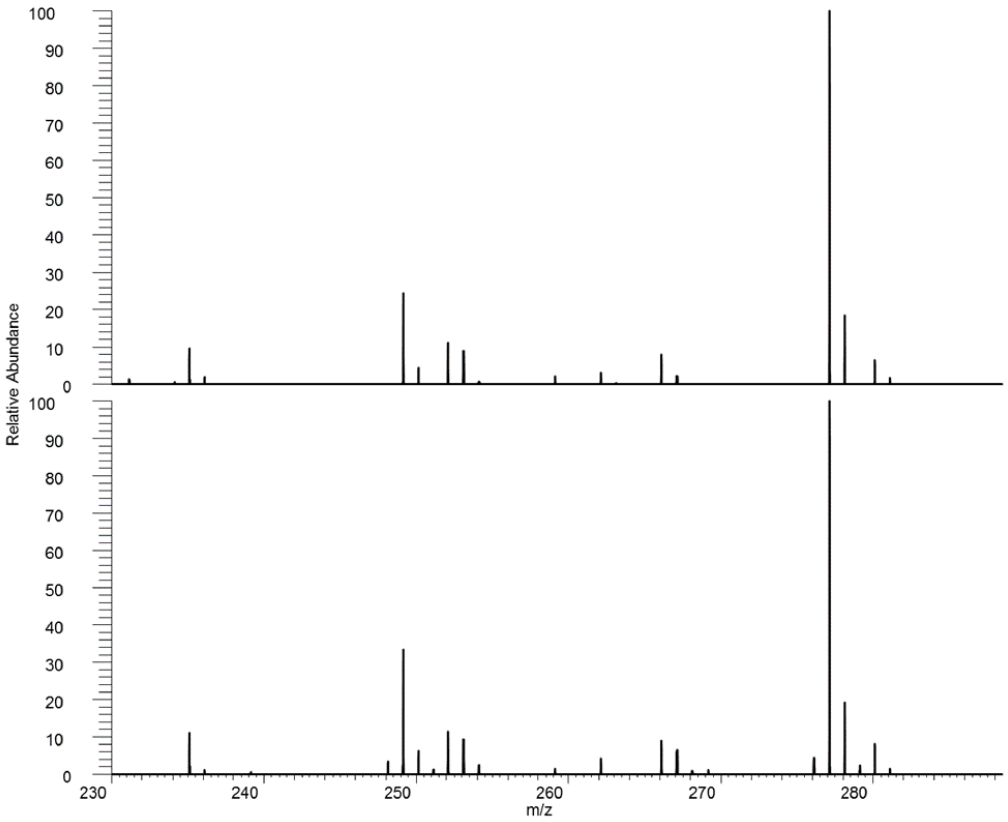


Figure S28. Fragmentation Patterns of Sugiol (HCD = 30). Fragmentation patterns of the purified compound (top) match fragmentation patterns of the compound found within the *S. miltiorrhiza* mixture (bottom).

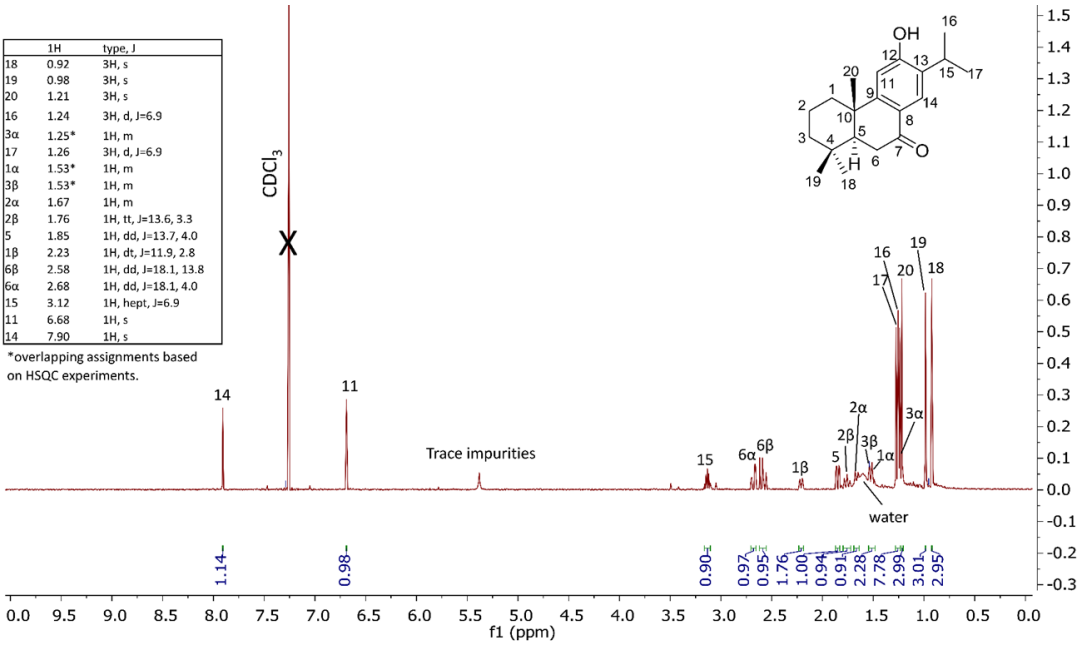


Figure S29. ^{1}H NMR Data for Sugiol (500 MHz, $CDCl_3$).

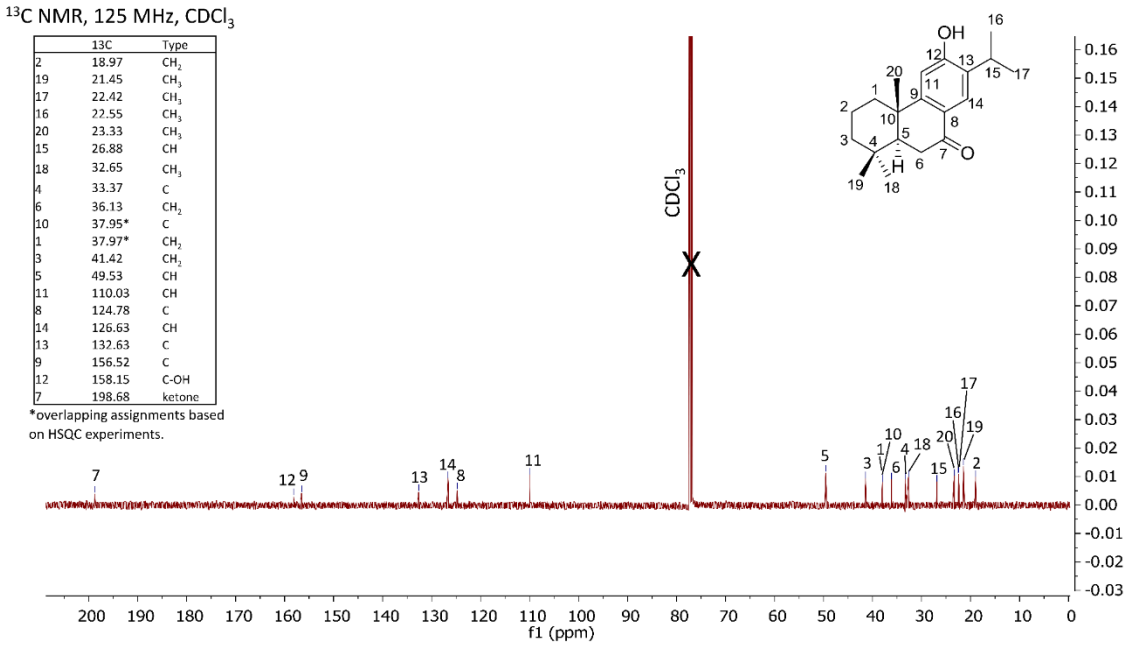


Figure S30. ^{13}C NMR Data for Sugiol (125 MHz, $CDCl_3$).

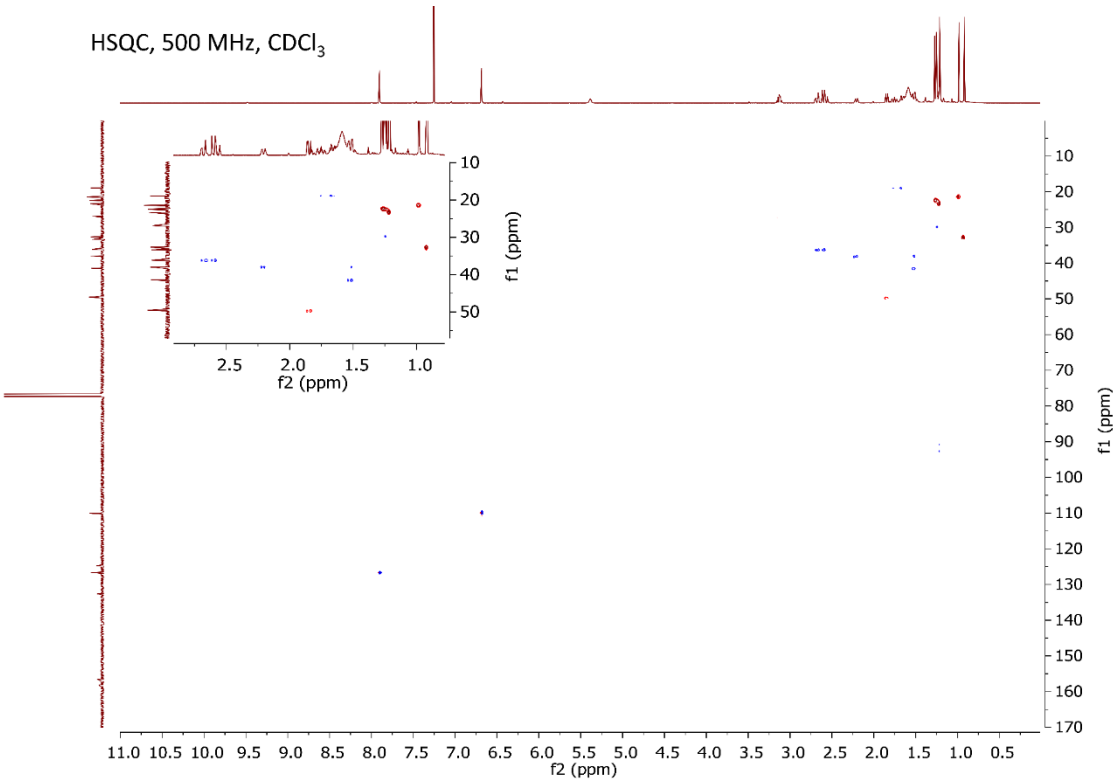


Figure S31. HSQC Data for Sugiol (500 MHz, $CDCl_3$).

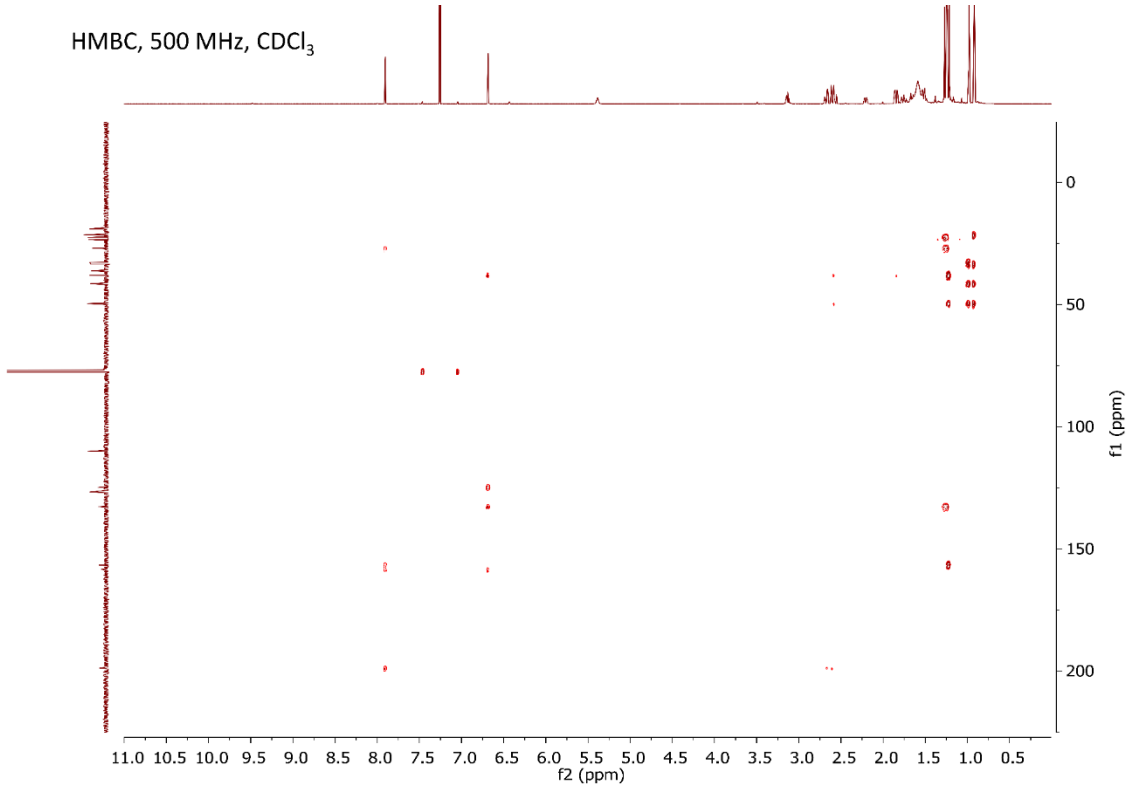


Figure S32. HMBC Data for Sugiol (500 MHz, $CDCl_3$).

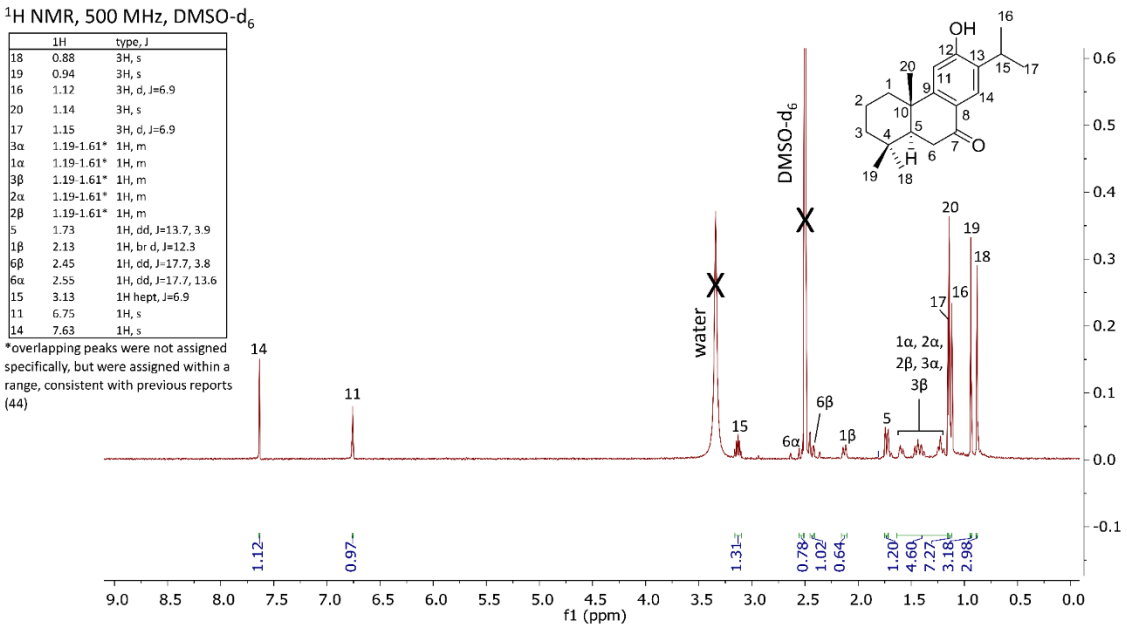


Figure S33. ^{1}H NMR Data for Sugiol (500 MHz, DMSO-d$_6$). Data are consistent with previous reports (322).

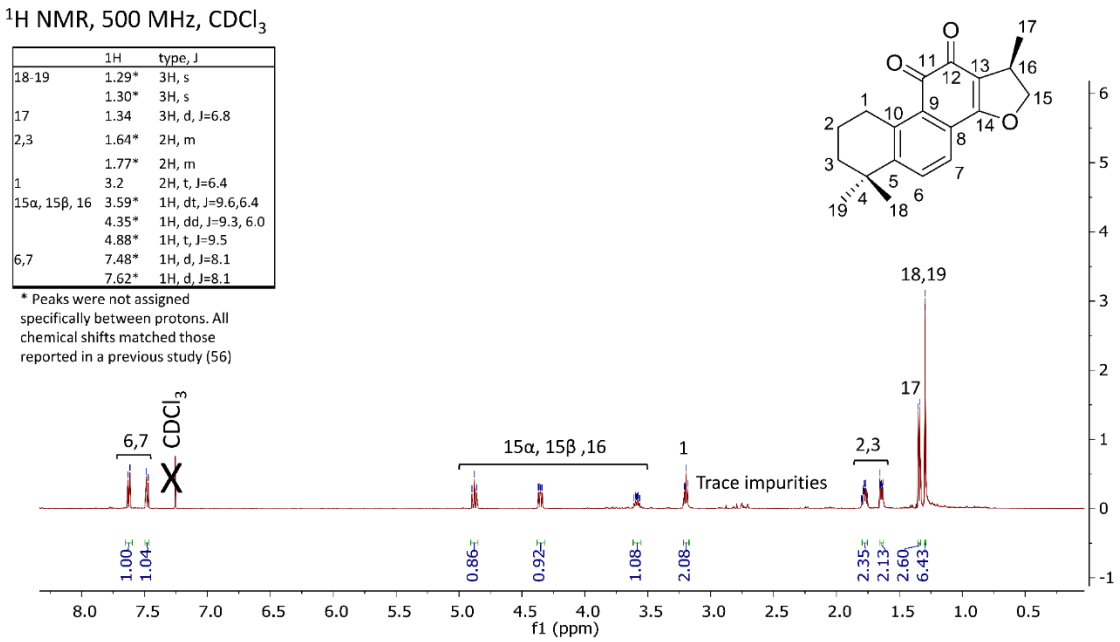


Figure S34. ^{1}H NMR Data for Cryptotanshinone (500 MHz, $CDCl_3$).

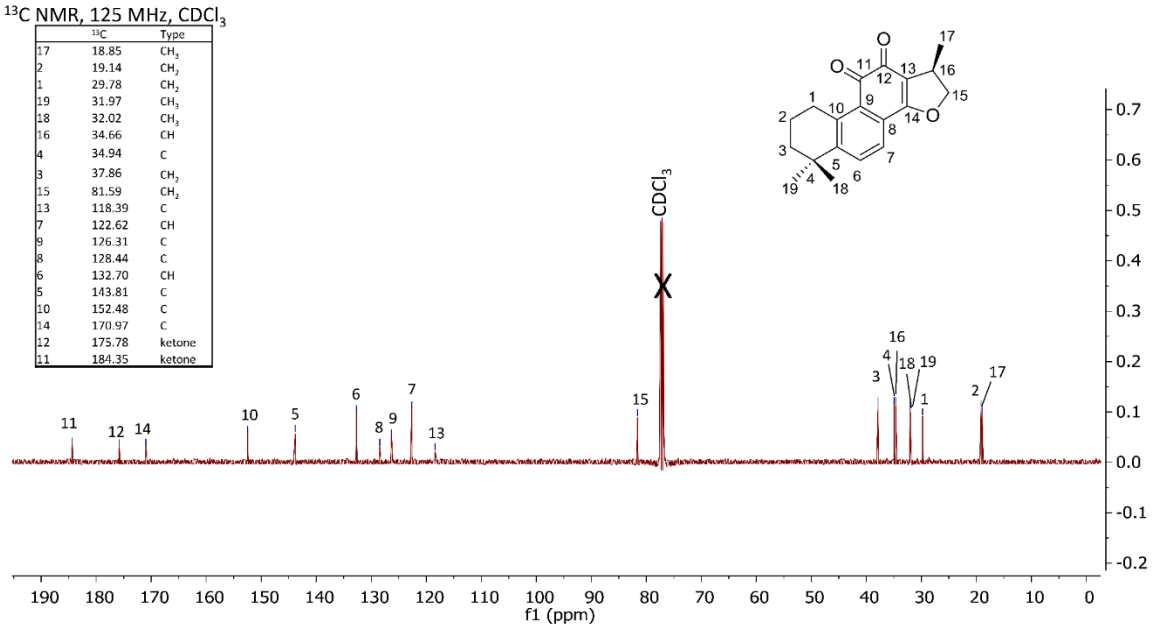


Figure S35. ^{13}C NMR Data for Cryptotanshinone (125 MHz, $CDCl_3$). Traces are consistent with previous reports (333).

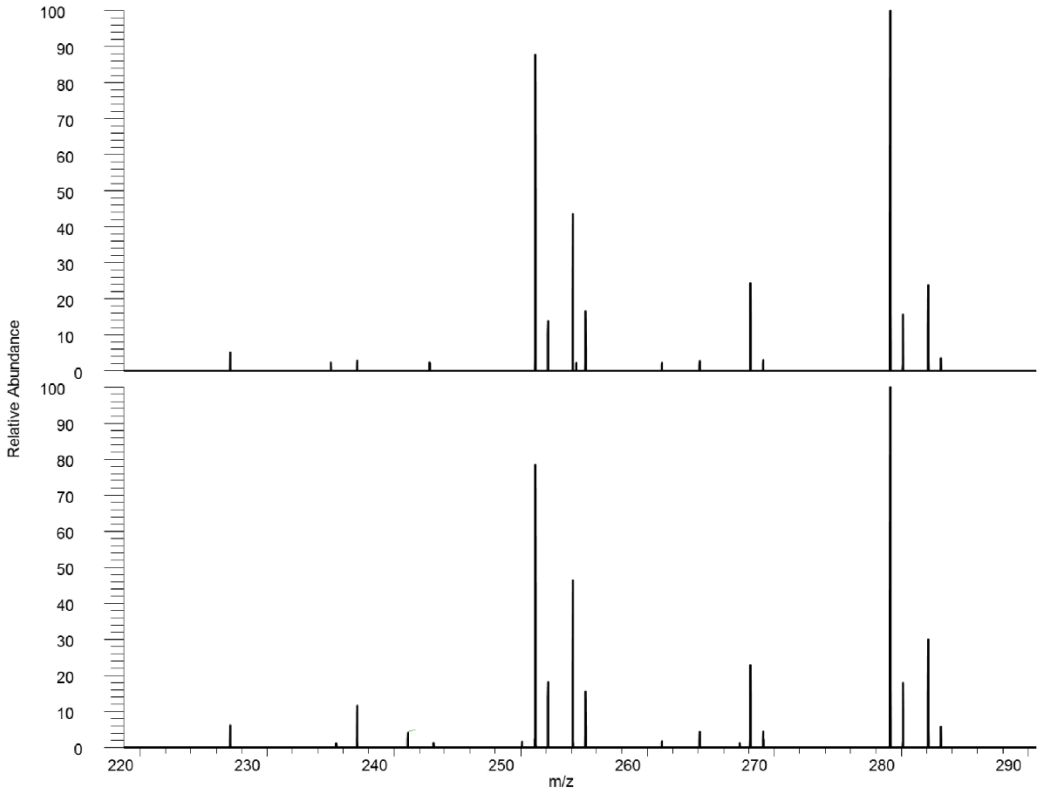


Figure S36. Fragmentation Patterns of Cryptotanshinone (HCD = 30). Fragmentation patterns of the pure standard compound (top) match fragmentation patterns of the compound isolated from the *S. miltiorrhiza* mixture (bottom).

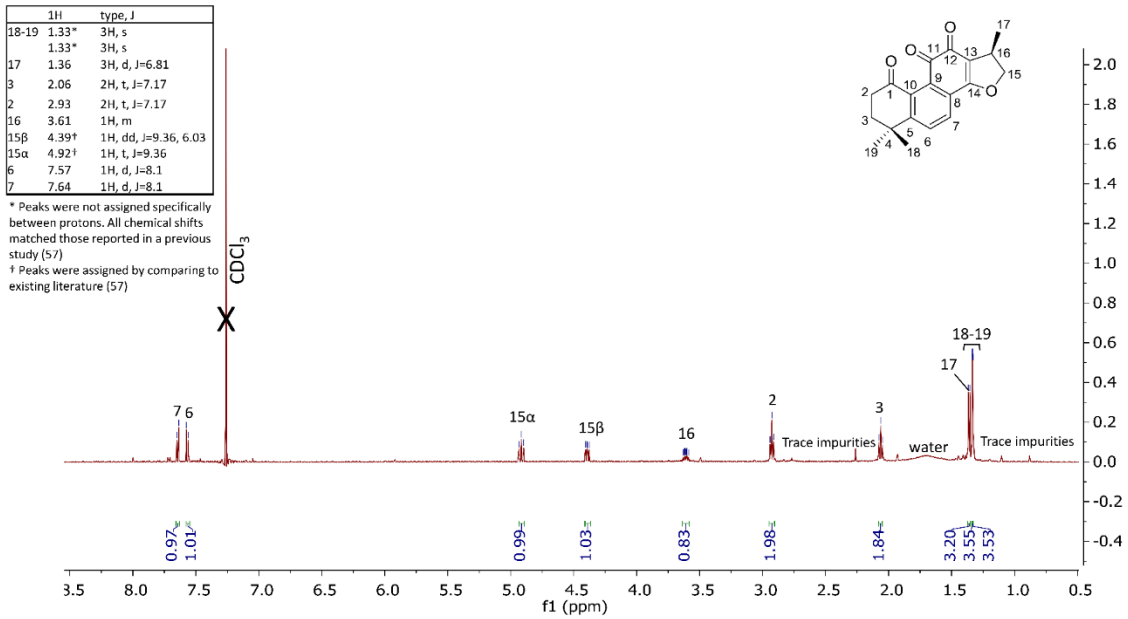


Figure S37. ^{1}H NMR Data for 1-oxocryptotanshinone (500 MHz, $CDCl_3$).

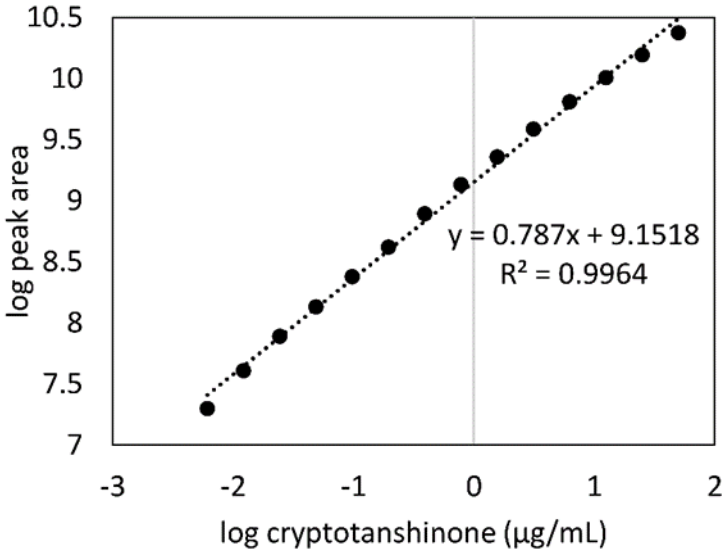


Figure S38. Calibration Curve of Cryptotanshinone used to Quantify Cryptotanshinone in each *S. miltiorrhiza* Fraction.